Transactions of the ASME

FREQUENCY-RESPONSE SYMPOSIUM

IRD Frequency-Response Symposium—Foreword 1145

The Regeneration Theory . . . . . *H. Nyquist* 1151

Early History of the Frequency-Response Field . . . . . *A. C. Hall* 1153

Frequency-Response Data Presentation, Standards and Design Criteria . . . *Rufus Oldenburger* 1155

Sine-Wave Generators . . . . . *D. W. St. Clair, L. W. Erath, and S. L. Gillespie* 1177

A Bibliography of the Frequency-Response Method as Applied to Automatic-Feedback-Control Systems . . . . . *A. M. Poole* 1185

The Frequency-Response Approach to the Design of a Mechanical Servo . . . . *H. A. Helm* 1195

Determination of Transient Response From Frequency Response . . . . . *A. Leonhard* 1215

Analysis of Regulating Systems With Particular Reference to Speed Control . . *R. H. Macmillan* 1237

Application of Frequency-Analysis Techniques to Hydraulic Control Systems . . . *A. C. Hall* 1245

Synthesis of Optimum Feedback Systems Satisfying a Power Limitation. . . . *J. H. Westcott* 1253

A Uniform Approach to the Optimum Adjustment of Control Loops . . . . . *R. C. Oldenburg and Hans Sartorius* 1265

Recent Advances in Nonlinear Servo Theory . . . . . *J. M. Loeb* 1281

A Statistical Approach to Servomechanisms and Regulators . . . . . *M. J. Pélegrin* 1291

Control-System Behavior Expressed as a Deviation Ratio . . . . . *J. M. L. Janssen* 1303

Frequency-Response Analysis and Controllability of a Chemical Plant . . . . *A. R. Aikman* 1313

Frequency-Response Methods Applied to the Study of Turbine Regulation in the Swedish Power System . . . . . *V. Ojo* 1325

The Use of Zeros and Poles for Frequency Response or Transient Response . . . *W. R. Evans* 1335

Approximate Frequency-Response Methods for Representing Saturation and Dead Band . . . . . *Harold Chestnut* 1345

Stability Characteristics of Closed-Loop Systems With Dead Band . . . . . *C. H. Thomas* 1365

Experimental Flight Methods for Evaluating Frequency-Response Characteristics of Aircraft . . . . . *G. A. Smith and W. C. Triplett* 1383

TRANSACTIONS OF THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS

VOLUME 76 NOVEMBER 1954 NUMBER 8

Transactions

of The American Society of Mechanical Engineers

Published on the tenth of every month, except March, June, September, and December

---

OFFICERS OF THE SOCIETY:

Lewis K. Sillcox, *President*

Joseph L. Koss, *Treasurer* — C. E. Davies, *Secretary*

Edgar J. Kates, *Asst Treasurer*

COMMITTEE ON PUBLICATIONS:

Paul T. Norton, Jr., *Chairman*

Otto de Lorenzi — W. E. Reaser

Colin Carmichael — Kerr Atkinson

Joseph Schmerler, Peter Wallace } *Junior Advisory Members*

George A. Stetson, *Editor* — K. W. Clendinning, *Managing Editor*

REGIONAL ADVISORY BOARD OF THE PUBLICATIONS COMMITTEE:

Richard L. Anthony—I
John de S. Coutinho—II
William N. Richards—III
Francis C. Smith—IV
H. M. Cather—V
J. Russell Parrish—VI
J. Kenneth Salisbury—VII
John H. Keyes—VIII

Published monthly by The American Society of Mechanical Engineers. Publication office at 20th and Northampton Streets, Easton, Pa. The editorial department is located at the headquarters of the Society, 29 West Thirty-Ninth Street, New York 18, N. Y. Cable address, "Dynamic," New York. Price $1.50 a copy, $12.00 a year for Transactions and the *Journal of Applied Mechanics*; to members and affiliates, $1.00 a copy, $6.00 a year. Add $1.50 for postage to all countries outside the United States, Canada, and Pan-American Union. Changes of address must be received at Society headquarters seven weeks before they are to be effective on the mailing list. Please send old as well as new address.... By-Laws: The Society shall not be responsible for statements or opinions advanced in papers or ... printed in its publications (B13, Par. 4).... Entered as second-class matter March 2, 1928, at the Post Office at Easton, Pa., under the Act of August 24, 1912....

Copyrighted, 1954, by The American Society of Mechanical Engineers. Reprints from this publication may be made on condition that full credit be given the Transactions of the ASME and the author, and that date of publication be stated.

IRD Frequency-Response Symposium

Foreword

THIS issue of the Transactions of The American Society of Mechanical Engineers is devoted to the papers and luncheon addresses of the Frequency-Response Symposium held in New York, N. Y., December 1 and 2, 1953, as part of the Annual Meeting of the Society, and two papers on frequency response given at the June, 1954, Pittsburgh, Pa., meeting of the Society. The Symposium was sponsored by the Dynamic Systems Committee of the Instruments and Regulators Division.

Automatic-Control Design

The frequency-response approach to the design of automatic controls is the dominant one today. This approach in a strict sense may be said to have started with a paper in 1932 by Dr. H. Nyquist, a physicist at the Bell Telephone Laboratories. In this paper Dr. Nyquist was concerned with the stability of feedback amplifiers. Mr. Harold S. Black, an engineer at the same laboratories, used the results of this paper in his work on negative feedback, published in 1934. This work can be said to have made possible the transcontinental telephone as well as modern radio and television. Dr. Hendrik Bode, mathematician at the Bell Telephone Laboratories, extended the results of Dr. Nyquist and obtained some simple design criteria that enabled even the technician without much mathematical training to decide in many cases whether or not a feedback system would be stable.

With the advent of World War II it became evident practically simultaneously to both a group under the direction of Prof. Gordon Brown at the Massachusetts Institute of Technology, and a group at the Bell Telephone Laboratories, that the contributions of Nyquist, Bode, and Black could be applied advantageously to automatic-control problems as well as to those of communication. This interest in frequency-response techniques was precipitated in part by the rigid military requirements for aiming antiaircraft guns by radar.

The Frequency-Response Approach

In the study of a physical device one is generally concerned with inputs and outputs to the device. As the inputs are varied the outputs are affected. The study of inputs and outputs is often reduced to that of the relation between one input and one output. The reader is not to infer that inputs and outputs are necessarily places where material goes into and out of the system. Inputs and outputs are simply physical variables. Thus throttle position and car speed may be taken to be, respectively, the input and output of an automobile.

In a broad sense the frequency-response approach to the design of a physical device may be said to be that in which use is made of the response of the output of the device to sinusoidal oscillations of the input. The Frequency-Response Symposium was devoted to frequency response in this broad sense, especially as it applies to feedback systems. In this general sense frequency response was employed long before Nyquist and Bode. The use of frequency response in the design of electrical and acoustical filters is classical.

Before World War II frequency-response runs on aircraft instruments had been made by C. S. Draper and others. Under the supervision of Gordon Brown at the Massachusetts Institute of Technology, sinusoidal runs were made in 1940 on hydraulic transmissions to study stroking forces. This work led rather naturally to the use of Nyquist and Bode design criteria.

Before frequency-response analysis came into extensive use transient responses to step changes were employed to check or obtain the dynamical characteristics of physical systems. It can be shown (see Paper 54—SA-3 by Smith and Triplett) from the theory of Fourier transforms that step changes are heavily weighted in favor of low frequencies. Modern control applications often, if not generally, require response that is as fast as possible. The responses of the components to relatively high frequencies then become very important.

Importance of Frequency-Response Curves

The information obtained from frequency-response runs is often put in the form of frequency-response curves. These curves contain more information than can be put in differential equations. In writing down the differential equations for a physical system one can take into account only factors considered dominant and neglect the rest. Except for errors in measurement, the experimental frequency-response curves for a physical device truly represent this device. No two physical systems have the same frequency-response curves. Such curves are in a sense as characteristic of a physical system as fingerprints are of a human being. They are really differential equations in pictorial form.

The work of Nyquist and his associates enabled the designer, on the basis of frequency-response curves only, to form conclusions about the stability of physical systems. So many contributors were involved in the early applications to automatic controls of the Nyquist and Bode work on the stability of feedback systems that it is difficult to give a precise history. To corroborate the information the writer collected on the subject he personally interviewed H. Nyquist, H. Bode, H. S. Black, H. B. Farrell, and L. A. MacColl of the Bell Telephone Laboratories, N. B. Nichols of the Raytheon Manufacturing Company, H. Harris of the Sperry Gyroscope Company, A. Hall of the Bendix Research Laboratories, Profs. G. Brown, C. S. Draper, and D. Campbell of the Massachusetts Institute of Technology, as well as other pioneers in the study of stability of feedback systems by frequency response.

In 1941 Mr. Herbert Harris published a United States Government report (OSRD 454) in which he applied Nyquist's work to servomechanism design. In 1943 the Technology Press published Dr. Albert Hall's doctor's thesis, which contained a comprehensive study of the application of Nyquist's theory to servomechanisms. In 1944 Drs. Rudolf C. Oldenbourg and Hans Sartorius published their book on automatic-control dynamics from the frequency-response point of view. Subsequently, at least two-hundred papers and books have appeared on the subject of frequency response.

The book by Oldenbourg and Sartorius was translated into English by Prof. H. L. Mason and was published by the ASME in 1948.

In 1943, N. B. Nichols, at that time at the Radiation Laboratory at the Massachusetts Institute of Technology, visited E. B. Farrell and Axel Jensen of the Bell Telephone Laboratories, who were using logarithmic charts. As a result of this visit Mr. Nichols began employing Black and other logarithmic plots. In one of the books of the Radiation Laboratory Series, published in 1947, Nichols, W. P. Manger, and E. H. Krohn gave an exposition of the logarithmic approach to servo design. The book in 1948 by G. Brown and D. Campbell on servomechanisms was

a significant factor in educating students in frequency-response techniques and bringing these techniques to the attention of industry.

Frequency Response in Process Control

Although frequency response has been in general use in the servomechanism, aeronautical, and electrical fields for several years, it is only now that it is making an extensive and determined invasion of the process-control field and certain other areas of industry. This invasion precipitated the formation of the Dynamic Systems Committee, which was assigned the task of recommending standards for the presentation of frequency-response data. In recognition of the importance of the frequency-response field the Frequency-Response Symposium was organized.

Drs. Nyquist, Bode, Hall, Mr. Harris, and other early contributors to the theory of design by frequency response have since left this domain for other areas of activity (except that Hall is still concerned with applications). Although theoretical work on frequency response still continues in the United States, a considerable proportion of this activity goes on in Europe. For this reason active European as well as American engineering scientists in the frequency-response domain were invited to participate in the Symposium. The result was a program with experts from the United States, England, Holland, France, Germany, and Sweden. Although excellent automatic-control work is going on in Russia, especially in the nonlinear field, that country, unfortunately, was not represented at the Symposium. In the 1930's Dr. W. Cauer[1] in Germany and A. W. Michailow[2] in Russia already were working on frequency-response design techniques.

In addition to the generous help received from reviewers, before the papers of this issue were approved for publication the writer tried, by reading the papers himself and reworking them with the authors where it appeared desirable, to get the papers to satisfy certain standards of exposition so that the reading of the papers would not be unduly burdensome. Owing to limited time and the difficulty of transoceanic correspondence it was not possible to meet this objective entirely. The writer therefore has requested the authors to treat in their closures the points that require clarification.

Scope of the Symposium

The papers in this issue of the Transactions are intended to cover a wide range of interests, from those of the beginner who may have little mathematical training, to that of the expert on frequency response. The first three papers are all directed to the beginner as well as to the expert. The paper by Helm also is intended for the beginner, but an acquaintance with transfer functions and other mathematical background is assumed. The paper by the writer gives the recommendations of the Dynamic Systems Committee for the presentation of frequency-response data. The Committee hopes that the general adoption of these recommendations will facilitate the exchange of frequency-response information among the workers in the field.

To enable the maximum number of people to use the committee's recommendations the writer attempted to carry the beginner without mathematical training as far as possible into the theory of the frequency-response field. The paper gives a survey of design criteria which will enable the reader at least to make a start on the scientific design of automatic controls. The frequency-response literature is full of unproved statements. The writer has worked through the necessary mathematics and endeavored to give the expert as well as beginner a mathematically rigorous picture of the material covered.

The paper by D. W. St. Clair, L. W. Erath, and S. L. Gillespie gives a survey of the pneumatic, electrical, and mechanical-hydraulic sine-wave generators in use. By means of these generators the input to a physical device is oscillated so that experimental frequency-response curves of the device can be obtained. Nearly all of these oscillators are laboratory models and are not commercially available. The use of such oscillators is absolutely essential to the application of frequency-response methods. The committee believes that this paper will be particularly valuable to those individuals and organizations starting frequency-response work.

Bibliography on Frequency Response

The bibliography paper by Abraham M. Fuchs sheds light on the history of frequency response and the areas of application. The bibliography by Mr. Fuchs lists not only papers devoted entirely to frequency response, but also papers concerned with other subjects but containing some frequency-response information. Often papers devoted to other subjects contain some frequency-response curves of physical systems. An effort has been made to include such papers although the listing is not claimed to be complete. Responses often can be characterized by transfer functions so that papers on transfer functions may be considered to belong to the frequency-response domain.

Frequency-Response Applications

The paper by Mr. H. A. Helm is a down-to-earth treatment of the design of an automatic contour-following lathe by frequency-response methods. Mr. Helm shows how the proper insertion of physical components may be employed to improve vastly the stability of the lathe. This paper was presented at the June, 1954, Pittsburgh Meeting of the Society.

The papers by Mr. R. H. Macmillan, Dr. A. C. Hall, Mr. A. R. Aikman, and Mr. Vladimir Oja are largely devoted to applications. The paper by Mr. Macmillan gives a survey of methods of analyzing automatic-control systems and discusses the transfer functions in the problem of governing the speed of a gas turbine. This paper is directed to the educator as well as to the control engineer and scientist.

Dr. Hall treats hydraulic servosystems with electrical control components. Of particular importance is the transfer function he gives for an ordinary servo composed of a four-way valve and piston-cylinder actuator. An electrical analog is given for the nonelectrical components so that the entire problem may be viewed electrically.

The paper by Mr. Aikman gives the results of applying frequency response to actual processes in chemical plants. In particular, he treats the automatic control of a spray drier and the automatic control of a heat exchanger. Mr. Aikman employs British terminology which differs in certain respects from that used in the United States and that recommended by the ASME.

The application of frequency-response methods is expected not only to affect the design of the controls themselves but also to affect the design of the process so that optimum results may be obtained from treating the process and control as a unit. Frequency response is ideally suited to showing how the various com-

[1] The writer interviewed Prof. R. M. Foster of the Brooklyn Polytechnic Institute, who is well acquainted with the work of Wilhelm Cauer. According to Professor Foster, Dr. Cauer worked on passive networks from 1926 to 1937 and thereafter on active networks. Like Bode, he was concerned with obtaining phase from amplitude response, and vice versa. In 1940 Dr. Cauer published a monograph on "The Poisson Integral and Applications to the Theory of Linear Networks—I," republished later by Edwards Bros., Ann Arbor, Mich. Cauer was killed when the Russians entered Berlin. The Bell Telephone Laboratories were quite concerned about volume II of the cited work, wanting to know whether or not Cauer had gone beyond the efforts of Bode and his associates. As far as we know, volume II has not been published.

[2] "Methods of Harmonic Analysis in Control Theory," by A. W. Michailow, *Avtomatika i Telemekhanika*, No. 3, 1938, pp. 27–81.

ponents of a system affect the performance of the whole. Mr. Aikman hopes, as does the Dynamic Systems Committee, that a complete library of experimental frequency-response curves eventually will be available to the control designer.

In his paper Mr. Oja gives the theoretical and experimental results of making frequency-response runs on the entire Swedish power system and discusses the problem of the regulation of this system by the prime-mover speed governors. This is the largest physical system on which frequency-response runs are known to have been made. The power output of one 50-kw machine was oscillated sinusoidally and the frequency of the system recorded, while all governors on the system were locked, and the system frequency was held by manual regulation.

Obtaining Transient From Frequency Response

In his paper of theoretical and practical interest, Prof. A. Leonhard discusses methods of obtaining transient response from frequency response. The reader will recall that in the transient method of analysis it is customary to obtain the response of the system to step changes in the input, which are frequently either the setting of the controlled variable or the load on the system. It appears that the calculation of these transients always will be an important factor in the design of feedback systems. The designers of electrical amplifiers who have been concerned primarily with frequency response recently have returned to transient response and have found it informative to give both frequency and transient-response curves for these amplifiers.

Relations between the frequency and transient responses in the form of integrals, and graphical methods of evaluating these integrals, have been known for some time. However, Dr. Leonhard gives some novel methods of making these calculations graphically and efficiently. His techniques are illustrated by three practical examples, one of which involves the important problem of computing the transfer function of a system represented by a partial differential equation. To accomplish this the Laplace transform of this equation is taken.

Optimum Control for Linear Systems

The papers by Dr. J. H. Westcott and Dr. Marc J. Pélegrin use the statistical approach to which much impetus was given by Prof. Norbert Wiener. The contributions of Dr. Westcott, Dr. Rudolf C. Oldenbourg, Dr. Hans Sartorius, and Mr. J. M. L. Janssen are concerned with the problem of optimum control for linear systems. For practical purposes the reader may assume that a system is linear if in sinusoidal analysis, whenever the frequency is held fixed, the amplitude of the output variable is directly proportional to that of the input.

Dr. Westcott assumes that the spectral density function of the input signal is given, as well as the transfer function of the fixed elements and the maximum rating of the power source. These are properties of the system and its control which are assigned to the designer and of which he has little if any choice. The problem is now to choose the transfer function of the rest of the circuit so that optimum results are obtained. There is some degree of freedom in the definition of the word "optimum." By this word Dr. Westcott means that the mean-square error in the value of the controlled variable should be a minimum for the response of the controlled variable to the statistically defined inputs. By variational methods from the calculus of variations the problem of optimizing a system for linear components is solved, at least theoretically. Beyond a certain point it is shown that increasing the complexity of the control does not improve the performance of the system.

Drs. Oldenbourg and Sartorius treat the optimization problem from a somewhat different point of view. They first consider transients for which the system is originally in equilibrium, and where the output eventually settles out to its initial value. For this case they treat the problem of minimizing the "linear-control area," that is, the area between the time axis and the curve of the deviation of the output from equilibrium. They also treat the problem of minimizing the "quadratic control area," namely, the area between the curve of the square of the output deviation and the time axis.

Where the transients do not die out before the next disturbance comes along, the authors minimize the quadratic mean output for periodic input disturbances that can be represented by Fourier series; that is, they minimize the average ordinate on the curve for the square of the deviation of the output from a base value. All plots are on Cartesian co-ordinates.

The authors show how the coefficients in the rational transfer functions representing linear controls can be chosen to make the proper quantity a minimum. A practical optimum which combines the advantages of the linear and quadratic measures of errors is given and applied to a practical example.

"Deviation Ratio"

The paper by Mr. Janssen is devoted to a discussion of the concept of deviation ratio. Excluding set-point changes, in the usual regulating problem the controlled variable of a physical system is to be kept constant, regardless of how the system is disturbed. For a sinusoidal disturbance, for example, in the load on a system, the deviation ratio is defined to be the deviation of the output of the physical system when a control is placed on the system divided by the deviation of this output when the system is without the control. The values of the deviation ratio plotted against frequency give a measure of how much the control helps to hold the output constant.

Dr. Westcott showed that the average value of the logarithm of the deviation ratio over the frequency range from zero to infinity is zero for linear systems. Improvement in one area of the frequency spectrum thus can be attained in linear studies only by sacrificing performance elsewhere in the frequency spectrum. Hence one cannot indefinitely improve the output by going to more expensive and elaborate linear controls.

The papers on optimum controls are particularly valuable in that they tell the designer how well he could do with unlimited means and how close he is to the optimum with a given design.

Application to Nonlinear Fields

The frequency-response approach was given considerable impetus by the extension of frequency-response methods to nonlinear fields. The papers by Dr. Julien M. Loeb, Mr. Harold Chestnut, and Mr. C. H. Thomas are concerned with this extension. Although the number of physical systems that can be treated, at least in rough analyses, by linear methods is vast, there are areas where these methods fail. This is particularly true of accurate analyses. In fact, there are areas where each physical system is nonlinear. Thus if the amplitude of an input sinusoidal oscillation to a physical system is small enough there will be no response in the output. When the input amplitude is large enough one crosses the threshold and the output responds but this response is not sinusoidal. As the input amplitude is increased still further one may reach an area of linear operation (for practical purposes) where the output is also sinusoidal and its amplitude is proportional to the input amplitude. If now the input amplitude is sufficiently increased one may reach a condition of saturation where the output is no longer sinusoidal. It has been shown[3] that where nonsinusoidal outputs result one can still apply frequency-response theory, such as that of Nyquist, provided the fundamental frequency component is employed.

---

[3] "On Some Non-Linear Phenomena in Regulatory Systems," by L. G. Goldfarb, *Avtomatika i Telemekhanika*, vol. 8, Sept.-Oct., 1947, pp. 349–383.

"A Frequency Response Method for Analyzing and Synthesizing Contactor Servo-Mechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, part II, 1950, pp. 270–283.

Sinusoidal Response

There are physical components for which the output is independent of frequency but depends on the amplitude of the input sinusoidal oscillation. Because of the importance of such components, it is suggested that the expression "frequency response" be replaced by "sinusoidal response." In his paper Dr. Loeb treats outputs of physical systems that are related in a rather general way to the inputs by functions of amplitude and frequency, as well as derivatives of these quantities with respect to time. The theory developed by Dr. Loeb involves the use of series expansions and is applied to contactor servos. Dr. Loeb is concerned with such problems as Coulomb friction and gear backlash.

The results of Chestnut and Thomas originally were intended to be given in a joint paper, but owing to the large body of material on hand, it was not possible to cover this in one paper. Their papers are concerned with such nonlinearities as saturation and dead band. Mr. Chestnut discusses theoretical and practical results and treats the describing function relating amplitude-dependent outputs to inputs, where the outputs are independent of frequency. With transfer-locus plots, Mr. Thomas treats the problem of instability and studies servos with gear trains involving appreciable backlash.

Statistical and Graphical Methods

The paper by Dr. Pélegrin, devoted to the statistical approach of Norbert Wiener, gives the results of some computer studies made in France. A particular servocircuit is computed by statistical methods and the autocorrelation function of an on-off relay is given. Classical and statistical methods are employed in the design of a variable inertia regulator to control the speed of an electric motor whose load varies.

In his paper Mr. Evans describes his root-locus method, rapidly coming into favor in the automatic-control field, and he shows how by graphical methods to obtain transient and frequency responses from his locus plots. He applies his methods to a load-damped servo.

Whereas the paper of Professor Leonhard is concerned with obtaining transient response from frequency response, the opposite problem is treated by Dr. G. A. Smith and Mr. W. C. Triplett. Making frequency-response runs on an aircraft is often time-consuming and expensive. For these and other reasons resort is often made to other types of disturbances, such as the unit step, pulse, and triangular inputs. The relative advantages of these disturbances are compared and methods are given for obtaining frequency response from the responses to these disturbances. The authors show how templates are used to fit a transfer function to an experimental frequency-response curve. Actual airplane differential equations are derived from basic laws of physics, and extrapolation for altitude and speed is treated.

The papers of this issue are intended to bring out the use of frequency response for checking differential equations, for diagnosing malfunctioning of equipment, and for the synthesis and direct design of automatic controls. The first two uses mentioned are probably the most prevalent ones.

Frequency-Response as Diagnostic Tool

The writer read the Harris 1941 and Hall 1943 pioneer frequency-response publications at the time that they came out but did not adopt frequency-response design methods because they were qualitative, and the writer had certain algebraic techniques for directly and rapidly obtaining the characteristic roots of linear systems. These roots gave much more quantitative information. On the other hand, the writer was practically forced to use frequency-response methods to solve the following problem, as well as to treat systems where the algebraic methods were too approximate.

When a gas-turbine speed governor on an aircraft gas turbine did not operate as calculated by the differential equations, direct analysis of the physics of the problem and transient runs on the turbine, where the trouble lay, did not reveal the cause of the difficulty. From frequency-response runs on the turbine made at the Woodward Governor Company in 1948, the writer was able to show that there was an unsuspected large dead time, subsequently found by the NACA Lewis Flight Propulsion Laboratory to be due to the time it takes for one combustion pattern to be formed from another. Nonlinearities and other phenomena, such as the impossibility of producing a perfect step, made it impossible to distinguish this dead time from the other phenomena until its existence was proved.

The solution to the troubles experienced in the design of the controls of the first American guided missile, the "Bat," were discovered by Dr. Hall by frequency-response methods. These and other examples emphasize an important phase of the frequency-response field. The checking of the differential equations of aircraft by airplanes in flight has revealed new factors in the design of aircraft. Extensive work in this direction has been done by Cornell Aeronautical Laboratories, Douglas Aircraft Company, and numerous other organizations.

In co-operation with the writer's company, one of the electrical power companies plans to use frequency response for routine testing of its steam and hydrogovernors as well as other components. The plan is to obtain the dynamic characteristics of the components when they are installed. After a suitable period of service, such as a year, frequency-response runs will be made on the same components.

If the curves have not changed, it means that these components are functioning correctly. If they have changed, an experienced and properly trained person often, if not always, can tell from the curves what has gone wrong so that repairs may be made. Frequency response is being considered by the writer's company for routine production tests of certain governors.

Since World War II over $2,000,000,000 worth of guided missiles have been designed with the aid of the Nyquist and Bode stability criteria. Several companies are employing frequency-response-stability theory in the design of instruments and automatic controls.

It is hoped that this issue of the Transactions will stimulate expansion of activities such as those outlined in the foregoing.

Acknowledgments

Much credit for the success of the Symposium must go to the authors, to the reviewers of the papers, and to the companies whose financial assistance made possible the participation by the foreign experts.

The papers were reviewed by the following: William R. Ahrendt, Frank Ainsworth, J. E. Bigelow, Gordon S. Brown, Page S. Buckley, William I. Caldwell, Donald Campbell, Gerald H. Cohen, Charles Concordia, Paul S. Creager, G. Forrest Drake, Donald P. Eckman, R. I. Edelman, Herbert Harris, S. P. Higgins, Jr., Ernest G. Holzmann, John A. Hrones, R. M. Hutchinson, Arthur S. Iberall, E. Calvin Johnson, R. W. Jones, Adam G. Kegel, Ralph J. Kochenburger, R. K. Koegler, Y. W. Lee, Sidney Lees, Mark Leum, R. S. Lovett, Julian E. McCarley, Harvey A. Miller, James Mozley, N. B. Nichols, William Owens, Henry Paynter, C. Desmond Pengelley, D. W. Pessen, Eugene Pike, David W. St. Clair, O. H. Schuck, R. C. Seamans, Jr., Charles J. Stukas, Curtis Thompson, William E. Vannah, G. A. Waldie, Elliot H. Woodhull, and Fred Woods.

Financial assistance was given to the foreign participants by the Bendix Aviation Corporation, Bristol Company, Carbide and Carbon Chemicals Company, E. I. du Pont de Nemours & Company, Foxboro Company, General Electric Company, Leeds and

Northrup Company, Manning, Maxwell and Moore, Inc., Minneapolis-Honeywell Regulator Company, Panellit, Inc., Raytheon Manufacturing Company, Royal-Dutch/Shell-Laboratory of Holland, Taylor Instrument Companies, and the Woodward Governor Company.

The writer wishes to express his appreciation to the staff of the Woodward Governor Company for its co-operation, in particular to Mrs. Helen Haarvei who undertook much of the secretarial work connected with the organization of the Symposium, to Mr. Mark Leum for his editing of the manuscripts, and to the writer's brilliant colleague, Mr. G. Forrest Drake, for his helpful suggestions.

Rufus Oldenburger
Chairman and Editor
Frequency-Response Symposium

The Regeneration Theory

By H. NYQUIST,[1] NEW YORK, N. Y.

The following is an explanation of the circumstances which led the author to write his 1932 paper on the "Regeneration Theory."[2]

FOR some years prior to the time of writing the paper[2] in question, the technical departments of the Bell System had been engaged in developing a carrier system for use in cables. Previously the long telephone circuits in cables had been confined to the audible frequency range and this limited the number of circuits obtainable to three circuits per quad of conductors. This yield was limited further by the need for four-wire operation on the longest circuits. Four-wire operation means that the two directions of transmission required separate conduction paths. The net result was that for these long circuits eight conductors or two quads were required to obtain three telephone circuits.

Multiplex Operation

The carrier art as practiced on open wire circuits was well advanced, and for such circuits it had resulted in substantial economy in conductor costs because a large range above the audible was used and this made multiplex operation possible. It was natural to think that similar carrier operation would be applicable to cable circuits as well. But cable operation posed a number of problems. Because the conductors were small the attenuation was large and this required close repeater spacing and special techniques for compensating for temperature effects on the attenuation. Because the conductors were close together there were new problems in coupling between circuits which had to be taken care of.

Of these various problems, we are here concerned with the repeater. Mr. Black proposed a negative feedback repeater and proved by tests that it possessed the advantages which he had predicted for it. In particular, its gain was constant to a high degree, and it was linear enough so that spurious signals caused by the interaction of the various channels could be kept within permissible limits.

For best results, the feedback factor, the quantity usually known as $\mu\beta$, had to be numerically much larger than unity. The possibility of stability with a feedback factor greater than unity was puzzling. Granted that the factor is negative it was not obvious how that would help. If the factor was minus 10, the effect of one round trip around the feedback loop is to change the magnitude of an original current from, say, 1 to −10. After a second trip around the loop the current becomes 100, and so forth. The totality looks much like a divergent series and it was not clear how such a succession of ever-increasing components could add to something finite and so stable as experience had shown.

Author's Part in the Problem

The missing part in this argument is that the numbers that describe the successive components 1, −10, 100, and so on, represent the steady state, whereas at any finite time many of the components have not yet reached the steady state and some of them, which are destined to become very large, have barely reached perceptible magnitude. My calculations were principally concerned with replacing the indefinite divergent series referred to by a series which gives the actual value attained at a specific time t. The series thus obtained is convergent instead of divergent and, moreover, converges to values in agreement with the experimental findings.

This explains how I came to undertake the work. It should perhaps be explained also how it comes to be so detailed. In the course of the calculations, the facts with which the term conditional stability have come to be associated, became apparent. One aspect of this is that it is possible to have a feedback loop which is stable and can be made unstable by increasing the loop loss. This seemed a very surprising result and appeared to require that all the steps be examined and set forth in full detail.

---

[1] Assistant Director of System Studies, Bell Telephone Laboratories, New York, N. Y.

[2] "The Regeneration Theory," by H. Nyquist, *Bell System Technical Journal*, vol. 11, 1932, pp. 126–147.

Luncheon address presented at the Frequency Response Symposium, held at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society.

Early History of the Frequency-Response Field

By A. C. HALL,[1] DETROIT, MICH.

THE Dynamic Systems Committee of the Instruments and Regulators Division of the Society should be congratulated for sponsoring this outstanding Frequency Response Symposium. Success has resulted only from the committee's devotion to organizing the detail for such stimulating sessions.

Early Impetus to Control Technology

Real progress results from strong stimulation. Upward of two decades ago the growing chemical industry supplied such stimulations for the process-control field and the resulting growth of the instrumentation industry is well known.

Similarly, a powerful impetus to the development of feedback-control technology began in 1940 as a result of the extreme technical problems arising from the second world war. Under the sponsorship of OSRD and the military, work was initiated at several institutions to find solutions to control problems of the most serious type. National security not only demands all that existing technology can provide but requires that conceivable advances in technology be made in incredibly short periods. Much of the work, of necessity, was done independently because of the restrictions of military security and, on the technical level, there was limited discussion of the problems and methods of attack. As a result some of the work was attacked in a parallel fashion, and even today with a perspective of more than a decade many questions on priority of accomplishments remain unsettled. I do not think it is appropriate or necessary to attempt to unravel the wartime activity. However, it is possible to see broad trends that occurred during the wartime period and to recognize the effects that these trends of activity had, and are still exerting, on the feedback-control field.

The war brought three major problems to the controls engineer. The first was handling problems and systems of considerable dynamics complexity dictated by the requirements of more accurate and rapid fire-control systems. The second problem was that of designing systems that would cope with large amounts of noise, occasioned by the use of radar in fire control. The third problem, raised by the guided missile, was that of designing so accurately a dynamic system that it could be used successfully almost at once with a negligible period of field trials.

Improving Fire Control

The first major problem faced in early 1940 was that of improving the accuracy of fire control for both ground and shipboard applications. The solution of this problem involved improvement of the performance of servo equipment for controlling the position of guns, the improvement of the accuracy of directors, and, later, in developing radar location equipment for fire-control purposes.

At the beginning of the war in Europe basic inadequacies lay, first in feedback-control components, second in systems concepts, but above all, inadequacies lay in our understanding of the problems and the methods of attack that would be effective. When the understanding of the problem is deepened to the point that fundamental limitations are grasped in spite of obscuring detail, the problem is well on the way toward being in hand. The invention of new systems will follow and attention can be focused on redesigning or designing around limiting components. Engineering efforts can be applied to the problem with the conviction that the solution will be significant. The role of analytical techniques primarily is to provide this understanding.

Analysis in Engineering Development

The primary responsibility of analysis in engineering development is frequently overlooked. Analysis should do more than provide a specific answer to a specific question raised for a particular system. Analysis should do more than generate proofs and generalizations for the satisfaction of the mathematician. The place of analysis in engineering development is to provide an understanding of the problem, an insight, a grasp, a visualization of the system operation for the engineer. It should set forth those factors that are basic, and limit fundamentally the results that can be obtained, and suggest what elements, by design, may be reduced to secondary position.

At the beginning of the war analysis techniques in use in the feedback-control field seemed sufficient for problems at hand. Fire-control requirements and systems were so complex, however, that the well-understood methods of transient analysis left much to the intuition of the engineer.

Problems of Radar Tracking

The second major problem of the war to be faced by the controls-systems engineer was that of radar tracking. Where the first problem, that of fire control, required the handling of very complex systems, the advent of radar required the controls engineer to design equipment to operate well in the presence of signals that he could not even describe in terms then in general use. The concept that a random process could be treated in terms of a frequency function was invaluable in the solution of a difficult and highly important problem. I well remember an instance in which M.I.T. and Sperry were co-operating on a control for an air-borne radar, one of the first such systems to be developed. Two of us had worked all day in the Garden City laboratories on Sunday, December 7, 1941, and consequently did not learn of the attack on Pearl Harbor until late in the evening. It had been a discouraging day for us because while we had designed a fine experimental system for test, we had missed completely the importance of noise with the result that the system's performance was characterized by large amounts of jitter and was entirely unsatisfactory. In attempting to find an answer to the problem we were led to make use of frequency-response techniques. Within three months we had a modified control system that was stable, had satisfactory transient response, and an order of magnitude less jitter. For me this experience was responsible for establishing a high level of confidence in frequency-response techniques.

[1] Technical Director Research Laboratories, Bendix Aviation Corporation.

Contributed by the Instruments and Regulators Division and presented at a luncheon meeting of the Frequency Response Symposium, held at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society.

Control of Guided Missiles

The third war-produced problem was the control of the guided missile. Here was introduced the requirement that the control design be so accurate that relatively few flights would be required to make field adjustments. Since the dynamics of the missile were as significant as any other part of the closed loop, there were no means of checking the system operation in the laboratory. In developing the air-to-ground missile, the "Bat," a challenge was provided by the need to consider a highly complex system, by the necessity of minimizing the noise introduced by the radar-tracking "homing head," and by the requirement that an operable missile be developed in a short time and with a limited expenditure of equipment in tests. The successful operation of the missile during tests in 1944 and 1945 was a direct result of the effective use of frequency-response-synthesis techniques.

The development was measurably aided also by the use of a crude "flight simulator" which permitted important checks on the over-all system operation to be made in the laboratory. The use of this tool improved our understanding of the system operation and enabled flight-test phenomena to be more easily related to control-system dynamics.

Progress Since 1945

Now, of course, the use of simulators or computers in solving dynamic and control problems is almost universal, and the advances in feedback systems since 1945 owes much to the effective use of such aids. Progress has been made in nonlinear systems, noncontinuous data systems, substantial improvement of components, and the application of controls to increasingly diverse problems. Our opportunities, particularly with the new computing tools at hand, are stimulating. However, we must not allow the availability of these tools or the seeming complexity of the problems to prevent us from gaining the real understanding or "feel" for the problems from which real progress is derived.

Frequency-Response Data Presentation, Standards and Design Criteria

By RUFUS OLDENBURGER,[1] ROCKFORD, ILL.

This paper gives standards for the presentation of frequency-response data. The standards are recommended by the ASME-IRD Dynamic Systems Committee to facilitate the exchange of information directly and through the medium of publications. The Committee recommends that magnitude curves be plotted on logarithmic co-ordinates, and phase on the linear scale versus frequency on the log scale. Further recommendations involve transfer functions, measurements, nonlinearity, ambient effects, and other factors. The paper was written with the technician in mind, as well as the automatic control engineer and scientist. For this reason the reader with little or no mathematical background is carried as far as possible into the theory of the frequency-response field. Basic design criteria, in common use, as well as others, are given, and it is shown that control design, in so far as dynamic performance is concerned, can be reduced, at least in rough analyses, to simple properties of curves that may be either theoretically or experimentally obtained. From the slope of one such curve the designer can often say whether or not a controller will be stable when placed on the system to be controlled. It is hoped that this paper will enable as many readers as possible to start on the scientific design of controls, and where they do so to use the standards here recommended. It is not implied that the reading of this paper is a substitute for a knowledge of some of the vast body of mathematics on which it is based. One objective of the paper is to give management, at least technical management, some insight into how automatic controls can be designed scientifically. The paper concludes with a discussion of the role the frequency-response approach plays among the design techniques available to the worker in the automatic control field.

Introduction

THE ASME-IRD Dynamic Systems Committee, hereafter referred to as the "DS-Committee," was formed by the IRD Executive Committee November 26, 1951, to recommend standards for the presentation of frequency-response data. The necessity for such standards became apparent simultaneously and independently to several members of the ASME who were employing frequency-response methods in the design of automatic controls. Some confusion had resulted because different companies presented the same frequency-response information in different ways. Thus one company would graph magnitude ratio versus frequency, and another the logarithm of the magnitude ratio versus the logarithm of the frequency. From a glance at the frequency-response curves for a physical component a great deal can be said about the dynamic properties of the component, provided that one is familiar with the co-ordinates used. In order that the curves of one company be readily interpreted by another it is highly desirable that they plot their curves on standard co-ordinates. These and other considerations led to the formation of the DS-Committee. Since its founding, this committee has made a thorough study of the matter of standards. The standards it wishes to recommend are incorporated in this paper.

To facilitate the understanding of the DS-Committee recommendations, and the frequency-response approach to control design, some of the background will be covered in this paper by the author, and design criteria given so that the technician with little or no mathematical training can with this information at least begin the scientific design of automatic controls. Most of the design criteria have appeared in various places in the servomechanism literature. One of the objectives of the paper is to give technical management some insight into how automatic controls can be designed scientifically.

Summaries of design criteria and committee recommendations are given near the end of the paper.

Recommendations for Magnitude and Phase Curves

Consider a physical system as in Fig. 1, with an input m and an

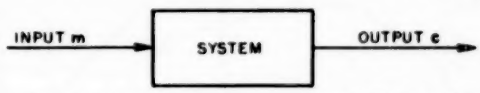


Fig. 1 Block Diagram of a Physical System With an Input m and Output c

output c. The input is a variable whose value m controls the value c of the output. An example is the case of a diesel engine where the throttle controls the speed of the engine. The variable m may be taken to be the throttle position in inches and c the speed of the engine in revolutions per minute. As is customary, the input and output are measured as deviations from equilibrium values, that is, values for which the engine speed and the throttle position are steady.

As the input is varied sinusoidally, so that $m = A \sin \omega t$ for the *magnitude* (amplitude) A, constant ω, and time t, the output eventually will vary sinusoidally if the system is what the mathematicians call linear. Experience and theory show that it is advisable for the reader to assume that the system he is studying can be treated as a linear system, unless proved otherwise.

The quantity ω is the product of 2π by the frequency f of the oscillation, that is, approximately $6.28\,f$. For steady oscillations the output of the linear system is $B \sin (\omega t + \phi)$ for a magnitude B and phase angle ϕ. For a linear system and a given frequency f the quantity B is proportional to the input magnitude A.

The quotient B/A is the *magnitude ratio* of the output to the input. In Fig. 2 the "magnitude frequency-response curve" of a diesel engine is plotted for the case of a load with small damping. For the example of Fig. 2 the quantity c is measured in units of a hundred rpm and the input m is in inches.

---

[1] Director of Research, Woodward Governor Company. Mem. ASME.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29-December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 3, 1953. Paper No. 53—A-11.

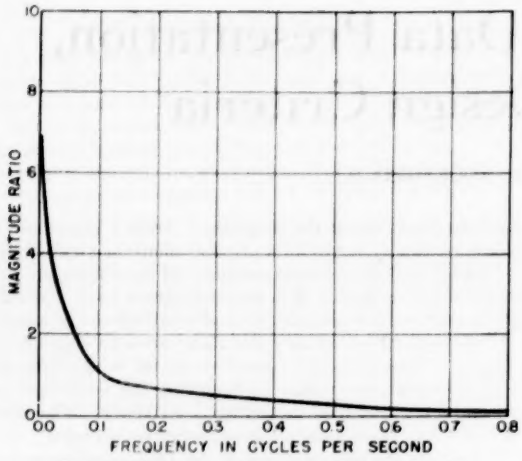


FIG. 2 MAGNITUDE CURVE OF A DIESEL ENGINE

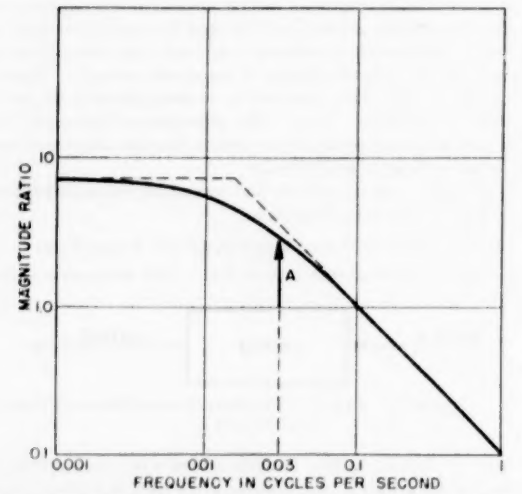


FIG. 3 LOGARITHMIC PLOT OF MAGNITUDE RESPONSE OF DIESEL ENGINE

The reader familiar with the frequency-response field will recognize the curve in Fig. 2 as that of a single-capacity system. Plotting the information of Fig. 2 on log co-ordinates,[1] one obtains the curve in Fig. 3. It will be noted that the curve in Fig. 3 can be approximated by two asymptotic straight-line rays as shown dashed in Fig. 3. The reader will recall that if the part of a straight line to one side of a point on the line is removed, what is left is a "ray." The curve formed by these rays is called the *asymptotic magnitude curve* of the engine. Since logarithms of real numbers to one base are proportional to the logarithms of these numbers to another base, the base associated with the log co-ordinates is immaterial.

Suppose that the engine speed is measured and that the difference between this speed and a speed setting, which it is desired to maintain, actuates a governor that positions the throttle. The difference can be considered to be the input to the governor and the throttle position to be the output. If the input is oscillated sinusoidally, the output will do the same, assuming again that the system is linear. A plot of a typical governor is given in Fig. 4.

[1] These plots are often called "Bode plots."

The curve in Fig. 4 can be approximated by the asymptotic curve shown dashed in Fig. 4, composed of two straight-line asymptotic rays represented by the first and last segments and two intermediate straight-line segments. To obtain this asymptotic curve it is really necessary to have the mathematical formula, that is "transfer function" for this curve. Where the technician is not acquainted with these functions, he will probably have to work with the actual response curves.

Suppose now that the governor is placed on the engine, but that the loop is open as shown in Fig. 5. Here the engine speed-measuring element is disconnected from the rest of the governor, theoretically if not actually. If the input e (actuating signal) to the governor is oscillated sinusoidally the output c (controlled variable) of the engine, namely, the engine speed (measured from equilibrium), will vary sinusoidally (for practical purposes). By mathematics it can be shown that the magnitude curve of the system composed of the governor and engine is obtained by merely adding ordinates on the magnitude curves for the components, plotted on logarithmic co-ordinates. Thus the vectors A and B in Figs. 3 and 4, at 0.03 cycle per second, are added to get the vector $A + B$ in Fig. 6 where the magnitude curve of the governor-engine open loop is plotted. It is convenient to think of the horizontal line labeled 1 as the horizontal axis because the logarithm of one is zero. The vectors A, B, and $A + B$ are thus measured from this axis.

To use the method of addition of ordinates it is necessary for the output of the controller and the input to the process (controlled system) to be measured in the same units. For the analysis of this paper the input e to the controller and output c of the process must be measured in the same units.

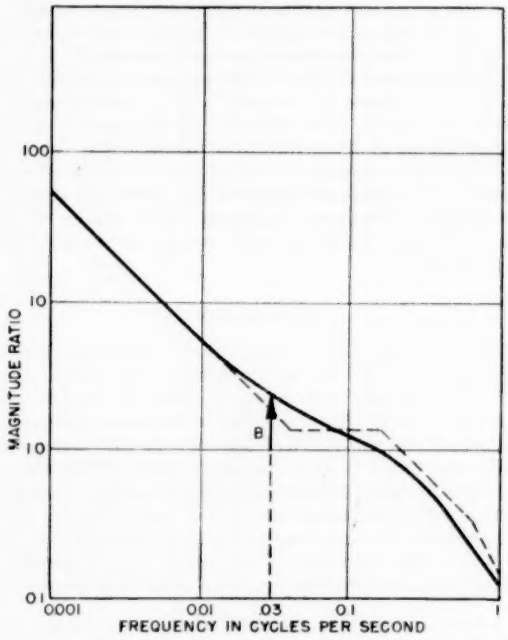


FIG. 4 MAGNITUDE CURVE OF AN ENGINE GOVERNOR

FIG. 5 MAGNITUDE CURVE FOR THE GOVERNOR-ENGINE OPEN LOOP

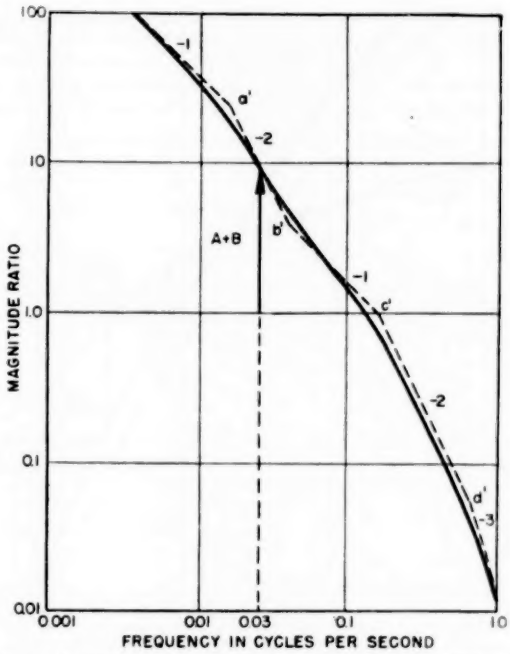


FIG. 6 MAGNITUDE CURVE FOR GOVERNOR-ENGINE OPEN LOOP

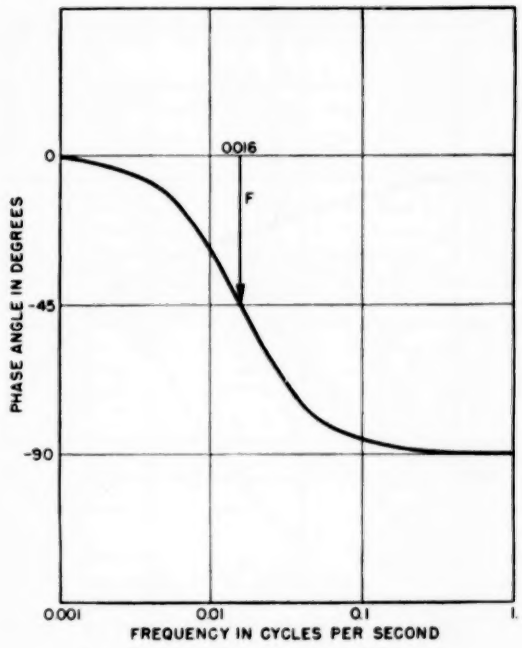


FIG. 7 PHASE CURVE FOR DIESEL ENGINE

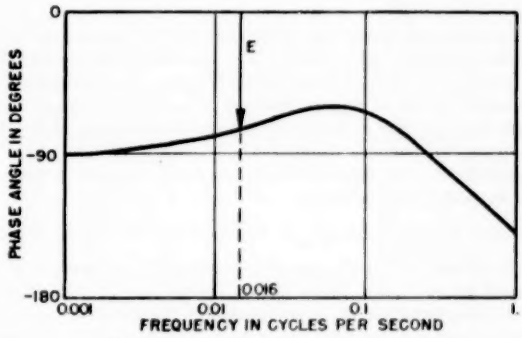


FIG. 8 PHASE CURVE FOR GOVERNOR

If the corresponding ordinates of the asymptotic curves in Figs. 3 and 4 are added, the broken line in Fig. 6 is obtained. This is the asymptotic magnitude curve of the governor-engine open loop. It is seen that the resultant curve obtained by using the straight-line approach is not much different from the correct curve.

Increasing the magnitude ratio of the governor-engine system by a factor K, regardless of frequency, is said to "raise the gain" of the system by K. The definition of "gain" is beyond the scope of this paper, and will be omitted (5).[1]

Raising the gain of the governor-engine system corresponds merely to shifting the magnitude curve of Fig. 6 upward. Gain is often readily adjustable in a physical system. When logarithmic co-ordinates are used one can often work with straight-line asymptotic approximations to the magnitude curves. Because of these considerations, the following recommendation is made by the DS-Committee:

Recommendation 1. It is recommended that magnitude curves be plotted on logarithmic co-ordinates with the "magnitude ratio" as the vertical co-ordinate and the "frequency" in cycles per unit time as the horizontal co-ordinate.

Except in England the frequency f is used in preference to the period P as the co-ordinate on the horizontal scale. Oscillators are generally marked for cycles per unit time, such as cycles per second, and it takes an additional computation to obtain radians ω per unit time. For these reasons the use of frequency is recommended.

It may be desirable to plot frequency on a nondimensional scale.

Let R be the magnitude ratio. It will be recalled that the magnitude ratio in *decibels* is $20 \log_{10} R$. The use of decibels is standard in the Bell Laboratories approach. However, the use of the logarithmic scale for magnitude ratio eliminates the need for converting magnitude-ratio values to decibels, thereby saving time in making plots of experimental data.

For the diesel engine the phase curve is shown in Fig. 7 with frequency and phase plotted on log and linear scales, respectively. The output lags the input as shown. The governor for the diesel engine has the phase curve of Fig. 8. Continued to the right the curve bends over and is asymptotic to the —180-deg line.

The phase angle of the governor-engine open-loop system at a frequency f is obtained by adding the phase angles of the components at the frequency f. Thus the vectors E and F in Figs. 8 and 7 add up to the vector $E + F$ in Fig. 9 where the phase curve is plotted for the system composed of the governor and the engine.

The use of a linear scale in angular degrees to express the phase difference between the input and output of a physical system is well established. The following is therefore recommended by the DS-Committee:

Recommendation 2. It is recommended that "phase" be plotted in degrees on a linear vertical scale and "frequency" in cycles per unit time on a horizontal logarithmic scale.

[1] Numbers in parentheses refer to the Bibliography at the end of the paper.

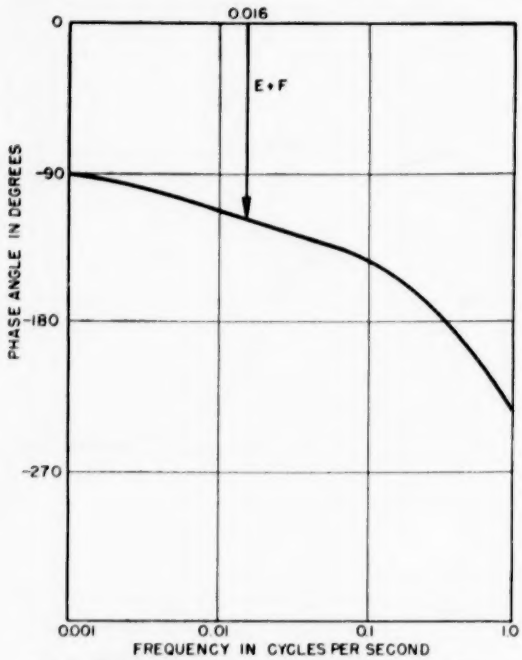


Fig. 9 Phase Curve for Governor-Engine Open Loop

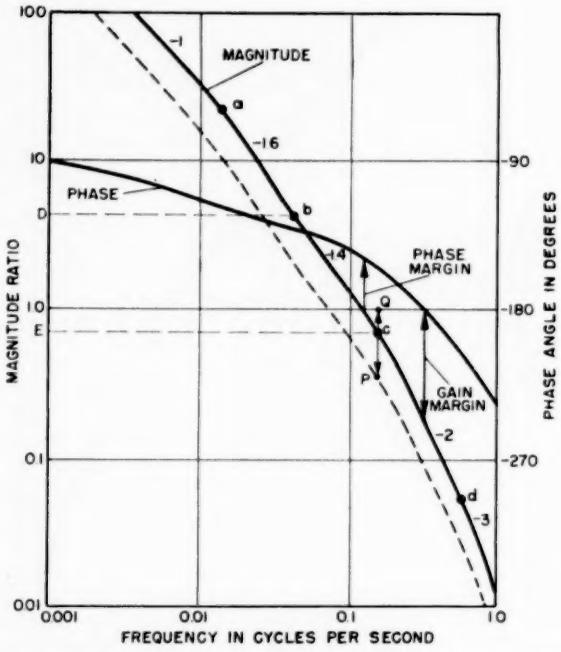


Fig. 11 Response Curves for Governor-Engine Open Loop

It will be convenient to use the same unit of measure for cycles on the magnitude and phase plots of Recommendations 1 and 2.

The magnitude and phase curves are just about as typical of a physical device as fingerprints are of a human being. These curves can be used to obtain the differential equations of physical systems, to diagnose troubles when there is disagreement between theory and experiment, and to design automatic controls for systems to be regulated.

Design Criteria Based on Magnitude and Phase Curves

From the magnitude and phase curves for the open loop the performance of the closed loop is determined (1). Now let r (reference input) be the speed setting in the governor, and c the actual engine speed (measured from equilibrium). The situation when the governor is regulating the engine speed is pictured in Fig. 10. A necessary requirement of the control is that the system of Fig. 10 be stable. From the magnitude and phase

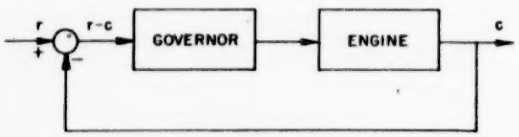


Fig. 10 Closed Loop

curves for the open loop of Fig. 5 a great deal can be said about the stability of the closed loop, and rough design rules can be stated. To save space it is convenient to plot both magnitude and phase curves on the same sheet of paper. For the engine example the curves are given in Fig. 11.

Dr. Harry Nyquist of the Bell Laboratories wrote the pioneering paper (2) of the frequency-response field. By advanced mathematics he solved the problem of stability for closed loops and linear systems by reducing the problem to the study of the open-loop response. Dr. Hendrik Bode of the Bell Laboratories extended this work and introduced (3) the concepts of gain and phase margin to be defined immediately. In terms of these concepts some simple design rules can be given.

Suppose that a small disturbance δ is applied to a system S, where δ eventually dies out. If for each such disturbance δ the resulting response of S dies out, the system S is said to be *stable*. Otherwise, S is *unstable*.

When the open-loop magnitude ratio is 1, we are at *gain crossover*. Suppose for the moment that there is only one point of gain crossover. This may be said to be the normal case. If we subtract the phase lag at gain crossover from 180 deg, the *phase margin* is obtained; that is, the phase margin is the sum of 180 deg and the phase angle. It is to be recalled that the phase angle is negative for a lag, and positive for a lead. If there is more than one point of gain crossover the angular difference just defined is formed for each point, and the minimum is defined to be the phase margin. The phase margin is indicated in Fig. 11 for the example under discussion. In this case the phase margin is about 30 deg. The reader will note that when the magnitude ratio (gain in the Bode sense) is 1, its logarithm is zero, and the magnitude curve crosses what may be conveniently thought of as the horizontal axis.

If we are at a condition where the phase lag is 180 deg (phase angle is -180 deg) we are at a point of *phase crossover*. Suppose first that there is only one such point. This may be considered to be the normal case. Let R be the magnitude ratio at phase crossover. The gain margin is then $1/R$ if R is less than 1, and equal to R if R is greater than or equal to 1. The gain margin is often measured in decibels. If there is more than one point of phase crossover, the "gain margin" is by definition the minimum of the quantities $\{R\}$ with $R \geq 1$ and $\{1/R\}$ with $R \leq 1$ for the R's at phase-crossover points. In Fig. 11 a vector has been drawn to indicate the gain margin. *It is not the length of the vector.* For the example of Fig. 11 the magnitude ratio R at

phase crossover is approximately 0.2. The gain margin is then 1/0.2, that is, 5. In decibels this is 20 $\log_{10}$ 5, or simply, 14 db.

The lag for the open loop may never reach 180 deg. In this case the gain margin is not defined. Similarly, the open-loop magnitude ratio may not reach 1, in which event the phase margin is not defined. Where a margin is not defined, and it is referred to in the design rules, the designer does not have to concern himself with it.

For the purpose of exposition it will be convenient to extend the definitions of gain and phase margins to some limiting cases. If

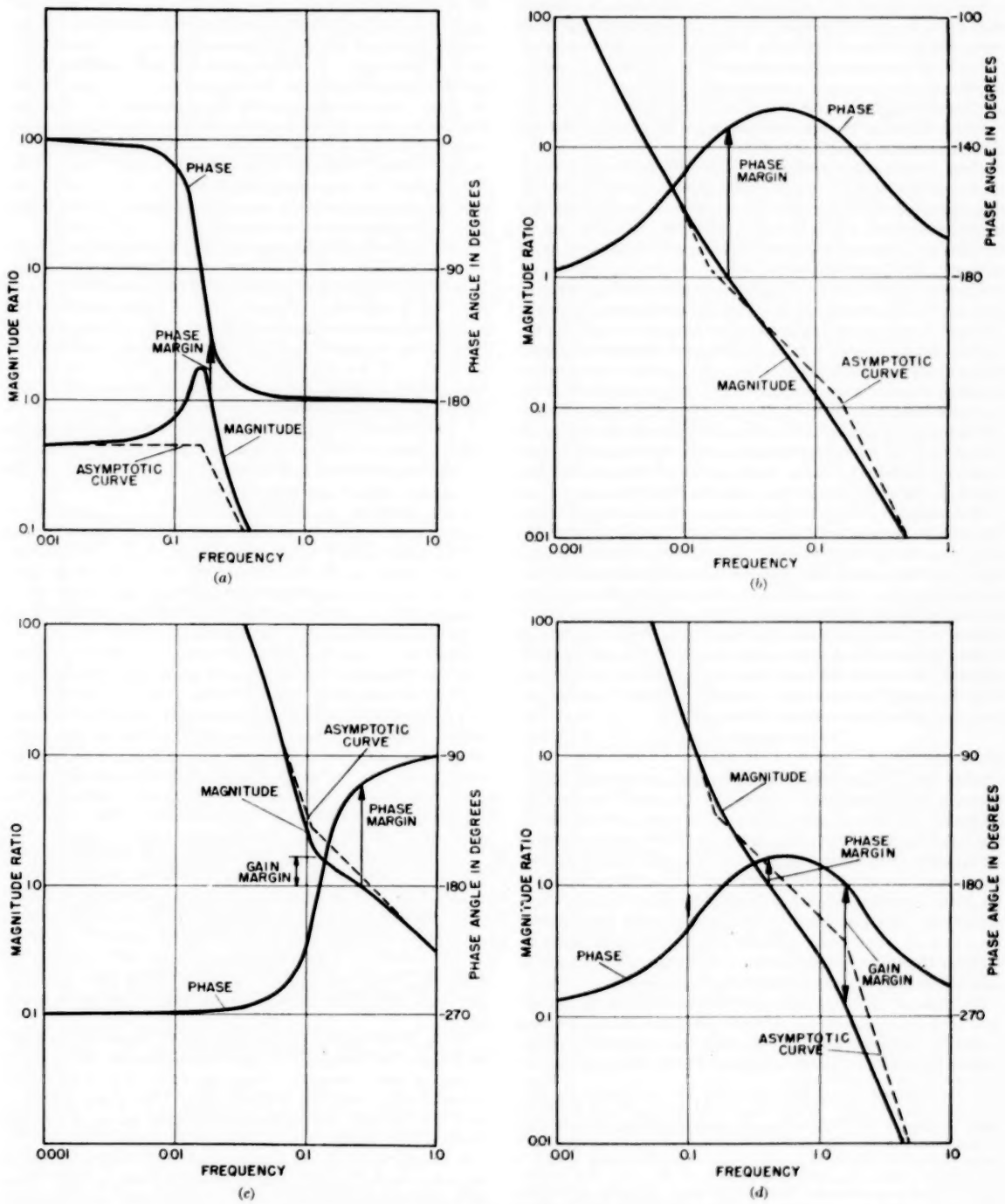


FIG. 12 GAIN AND PHASE MARGINS FOR SOME PHYSICAL SYSTEMS

the lag never reaches 180 deg but approaches 180 deg as the magnitude ratio goes to zero, the gain margin will be said to be infinite, and will then be denoted by ∞. Similarly, if the magnitude ratio never reaches 1 but approaches 1 as the lag goes to zero, the phase margin is said to be 180 deg.

From the laws of physics (4) and energy limitations it is known that as the frequency increases beyond all bounds the magnitude ratio of the open loop approaches zero or a finite positive number.

In Fig. 12 are shown gain and phase margins for some open-loop physical systems.

The following design rule is in common use (5–9):

Design Rule 1. The phase margin should be at least 30 deg and the gain margin at least 2.5 (8 decibels).

These rules can be justified on mathematical grounds. It can be shown by complex variable theory that the condition of 180-deg phase lag and magnitude ratio equal to 1 for the open loop is to be avoided. Rule 1 insures that this condition will be avoided by a definite margin. This will be clarified further in the discussion of transfer locus plots.

The foregoing design rule should be considered a rough criterion that enables the technician untrained in mathematics, as well as one who is trained, to take the frequency-response curves of the control and the system to be controlled, and eliminate some poor designs. It can be proved that, in general, it is not safe to allow smaller phase and gain margins. Thus a phase margin of 5 deg or gain margin of 1.05 would place the system in a sense so close to the border of instability that the solution would be highly oscillatory, and if the frequency-response curves were obtained all or in part from theory rather than experiment the neglected factors might throw the system into complete instability.

It can happen that the foregoing design rule is satisfied, but that the closed-loop system is still too oscillatory. An example will be given near the end of this paper. It can be shown, however, that for a wide range of physical systems the rules provide completely adequate control. The rules thus give a margin of safety that is often sufficient, but sometimes falls short. To be sure that a control will work, the technician without the background to go further into the theory might make use of a library of frequency-response curves that characterize good performance, and see to it that the curves he obtains for a specific problem are identical or close to a pair of curves known to be acceptable.

Design on Basis of Magnitude Curves Only

Bode (3) treated primarily physical systems that can be represented by ordinary linear differential equations with constant coefficients. Let a magnitude curve W be given and consider the systems of the foregoing type with the same W. Among the phase curves for these systems there is one for which the phase lags take on the smallest values. A system with such a curve is called a *minimum phase system.* For minimum phase systems Bode showed that the magnitude curve determines the phase curve.[4] Since it is sometimes more difficult to obtain the phase curve experimentally than the magnitude curve, it is desirable to have design criteria based on the magnitude curve alone.

From one of Bode's theorems (3) it follows that the phase angle at a given frequency f, for a minimum phase system, depends primarily on the slope of the magnitude curve (plotted on logarithmic co-ordinates) in the neighborhood of f and, if the slope is reasonably constant in the neighborhood of f, the phase angle is proportional to the slope of the magnitude curve. In fact, the phase angle is approximately $(90\ m)$ deg for the slope m.

To illustrate this relationship the reader is referred to the solid magnitude curve of Fig. 11. The part of the curve to the left of the point a has the slope -1 approximately. The reader will recall that a line with slope -1 falls one unit for each unit movement to the right; that is, the line makes an angle of 135 deg with the horizontal axis. By Bode's result the phase angle should be approximately -90 deg for low frequencies, which checks with the figure. For the segment c-d of the magnitude curve the slope is approximately -2 (the line drops two units for each unit movement to the right) and the corresponding phase angle should be about -180 deg, which also agrees with the figure. To the right of the point d the slope is approximately -3, whence the corresponding phase angle should be about -270 deg. For high frequencies the phase curve is asymptotic to the -270-deg line.

The slope of the curve from a to b is approximately -1.6, and from b to c about -1.4. Therefore we would expect the phase angle for the frequencies associated with the arc from a to c to be about -135 deg, which is correct. It is evident that there is a rough correlation between the slopes -1, -2, and -3 of the segments of the asymptotic curve of Fig. 6 and the actual phase angles.

For points on a long section of the magnitude curve, but away from the ends of this section where the slope m along the section is a fairly constant integer, the formula $-90\ m$ deg is a good approximation to the phase angle. Near the ends of the section the approximation becomes a very rough one. For a short section the approximation is also very rough.

If gain crossover occurs at a point interior to a long section of constant slope -2, the lag at gain crossover will be about 180 deg, and the phase margin will be numerically small. To insure an adequate phase margin the following rule (7) is used for the common case where the magnitude curve is falling at gain crossover:

Design Rule 2. The slope of the magnitude curve on logarithmic co-ordinates at and near gain crossover should be about -1.

The slope -1 has been chosen because -1 is definitely greater than the "dangerous" value -2, and at the same time it is generally, or at least often, desirable to stay as close to the slope -2 as is practical in order to obtain a transient response that is sufficiently oscillatory. Let f denote the frequency at gain crossover. *It is customary to design for a slope near -1 over the frequency range from $\frac{1}{2}f$ to $2f$.* The extent of the region over which the slope should be -1 to each side of the point of gain crossover depends on the slope away from gain crossover. With experience the designer will know about how much to allow for this.

The dashed curve of Fig. 11 is obtained by moving the solid magnitude curve of Fig. 11 down a distance equal to log 2. Design Rule 2 is satisfied by the dashed magnitude curve of Fig. 11 since the slope is approximately -1.4 (this is near enough -1 for practical purposes) at and near gain crossover. It is not so well satisfied by the solid magnitude curve since the slope on this curve rapidly changes to -2 as one passes gain crossover. The solid magnitude curve, however, still represents a stable system.

In Fig. 13 the magnitude and phase curves are shown for a stable system with the loop open, where the magnitude curve rises at a point of gain crossover. This may be considered to be the unusual rather than the common case. The phase margin is more than 180 deg, and the system has a large margin of stability when the loop is closed.

In plotting the frequency-response curves, it is not necessary to know the precise nature of these curves for all frequencies. Thus for a physical system with transients that endure 1 or 2 min it is not normally necessary to know the response at 100 cycles per

[4] In the process-control field dead time and distributed lag often play an important role. Where they do one does not have minimum phase systems, in fact, the ordinary linear differential-equations approach breaks down. The reader will then need both magnitude and phase curves. However, although in this case the design theory of the present section does not hold, the rest of the paper still applies. An extensive mathematical treatment of dead time and distributed lag has been given by Oldenbourg and Sartorius (10). Also see Bode (3), chapter 13.

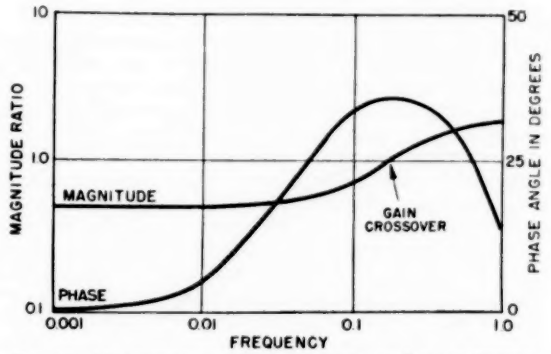


FIG. 13 MAGNITUDE AND PHASE CURVES OF STABLE SYSTEM

minute (cpm) or 0.01 cpm. In fact, a range of 10 cpm to 0.1 cpm is often sufficient.

The frequency-response curves should be plotted for an intermediate range of frequencies, and not necessarily for extremely low frequencies. Where gain and phase margins are defined these curves should be known for a range of frequencies that extend well beyond the frequencies at the crossover points. A common case is that where the asymptotic magnitude curve in the neighborhood of gain crossover is composed of three sections with slopes —2, —1, and —2, respectively, as shown in Fig. 14. The magnitude curve (solid or dashed) of Fig. 11 is an example to which Fig. 14 applies. To insure adequate stability margins the

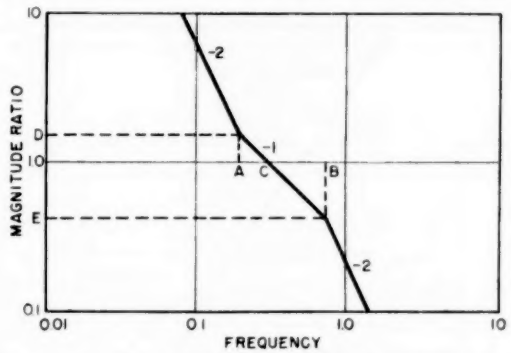


FIG. 14 PORTION OF COMMON ASYMPTOTIC MAGNITUDE CURVE

author recommends the following rule[5] based on a mathematical study of transfer functions (see Appendix to this paper for some mathematical details):

Design Rule 3. Consider a system with a portion of its open-loop magnitude curve approximated by the curve of Fig. 14 (composed of straight-line sections), where the sections with slope —2 are long compared to the middle section (section with gain crossover) with slope —1. For this system it is desirable to have the distance AC equal to about 1/3 to 1/2 of the distance AB. The quotient of the magnitude ratio associated with the point D by that for E should be at least 5 (ratio of frequency associated with B to that for A should be at least 5).

It can be easily shown that the magnitude ratio associated with the point D of Fig. 14, divided by the magnitude ratio associated with E, is equal to the frequency associated with B divided by the frequency associated with A.

It follows from Bode's phase-magnitude relationship that increasing the slope of a magnitude curve (making it less steep where the curve is falling) increases the phase angle. If now one or both sections adjoining the middle section with slope —1 are short sections with slope —2 adjoining sections with greater slope (not as steep), *the ratio of the frequencies* associated with the points B and A *can be somewhat less than* 5, such as 4.

For the governor-engine example of Fig. 11, the middle section bc of the solid magnitude curve has a slope about —1.4 while the adjoining sections ab and cd have the approximate slopes —1.6 and —2, respectively. The points a, b, c, and d marked on this figure are points of demarcation between sections with essentially different slopes, and may be called "break points." In the strict sense the expression *break point* applies to a point where straight-line sections of different slopes meet. For purposes of exposition we shall use the more liberal interpretation given here. On the magnitude curve in Fig. 11 the break points are not unique, but are chosen visually.[6] For the asymptotic magnitude curve they are unique. In Fig. 6 they are the points a', b', c', and d'. In a very rough way the section ab may be treated as one with a slope approximately equal to —2 and the section bc as one with the approximate slope —1. The vertical distance from c to b is the log 6.5, whence the quotient of the magnitude ratios corresponding to DE of Fig. 11 is 6.5. Design rule 3 is satisfied in this respect.

The sections $a'b'$, $b'c'$, and $c'd'$, of the asymptotic magnitude curve for the governor-engine example have the slopes —2, —1, and —2, respectively, Fig. 6. The vertical distance between b' and c' is log 4, whence the corresponding ratio is 4. This does not satisfy Design Rule 3. However, the section $a'b'$ is not long compared to $b'c'$, and the ray on the asymptotic curve to the left of a' has the slope —1. It follows that the *quotient of the magnitude* ratios associated with the points b' and c' *can be somewhat less* than the 5 of Design Rule 3, and the actual value 4 is quite satisfactory.

The point of gain crossover on the solid magnitude curve, Fig. 11, is near the right end of the middle section. For the asymptotic curve, Fig. 6, it is at the right end. Design Rule 3 is thus violated in this respect. Let the logarithms be taken to the base 10. Moving the magnitude curve down a distance 0.3, as shown in the dashed curve in Fig. 11, brings gain crossover at a point in agreement with Design Rule 3, namely, to or near the mid-point of the middle section. This is true whether we work with the magnitude curve or its asymptotic approximation.

As noted before, displacing the magnitude curve on logarithmic co-ordinates up or down *changes* the gain of the open loop accordingly. This is one of the major advantages of plotting magnitude on these co-ordinates. Doubling the gain merely multiplies each amplitude ratio by 2 and thus adds $\log_{10} 2$, that is, 0.3, to each ordinate. Halving the gain corresponds to moving the magnitude curve down a distance 0.3 since $\log_{10} 0.5 = -0.3$. As noted before, in the example of Fig. 11, we have halved the gain in going from the heavy to the dashed curve. *Changing the gain does not affect the phase curve.*

The foregoing application of Design Rule 3 may appear to lack precision, but this rule has been illustrated with an actual arbitrarily chosen example. Nevertheless, the rough methods given here yield good quantitative results.

The desire to have an adequate but not excessive phase margin is a reason for the foregoing rule. The author will introduce here an adjunct intended to help the designer secure adequate gain margin and stability. This adjunct is based on a study of transfer functions. It is assumed that we have the common case where the magnitude curve is falling at gain crossover.

Let R be the magnitude ratio at a point P. The *magnitude*

[5] In (9), p. 185, the ratio AC/AB equal to 1/3 is recommended.

[6] The writer asked six technicians one at a time to select the break points on this curve, given only the curve. All of them came remarkably close to the four break points of Fig. 11.

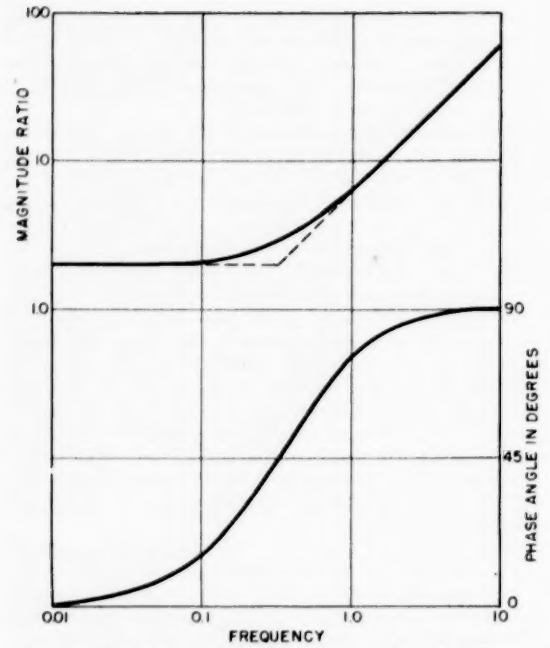


FIG. 15 FREQUENCY-RESPONSE CURVES OF LEAD COMPONENT

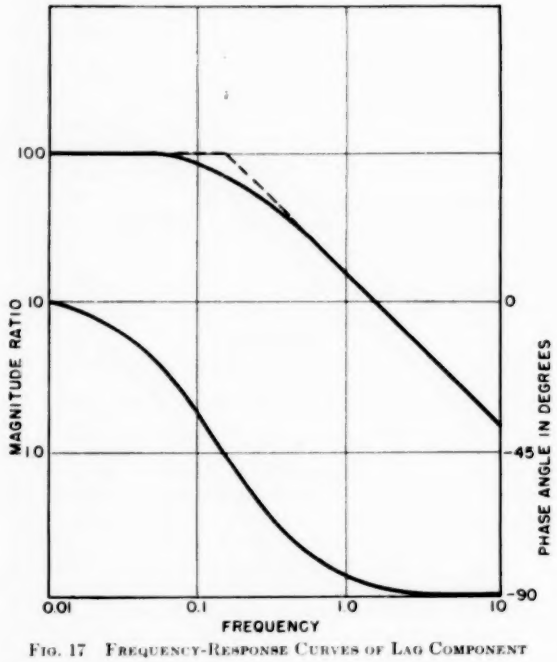


FIG. 17 FREQUENCY-RESPONSE CURVES OF LAG COMPONENT

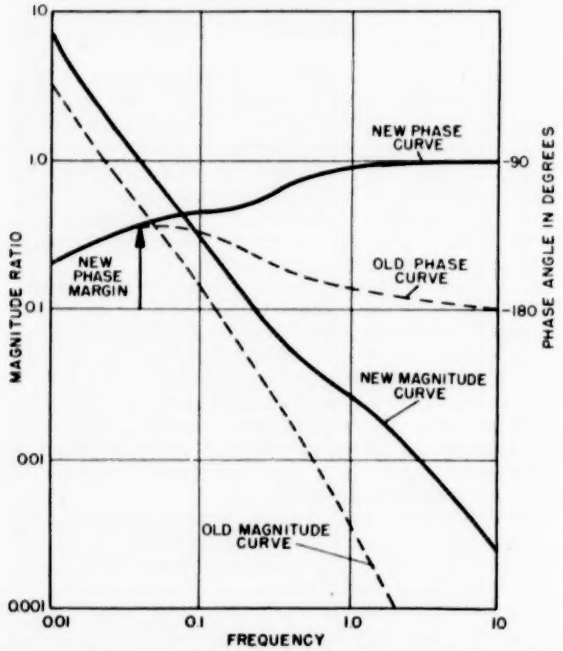


FIG. 16 FREQUENCY-RESPONSE CURVES MODIFIED BY LEAD COMPONENT

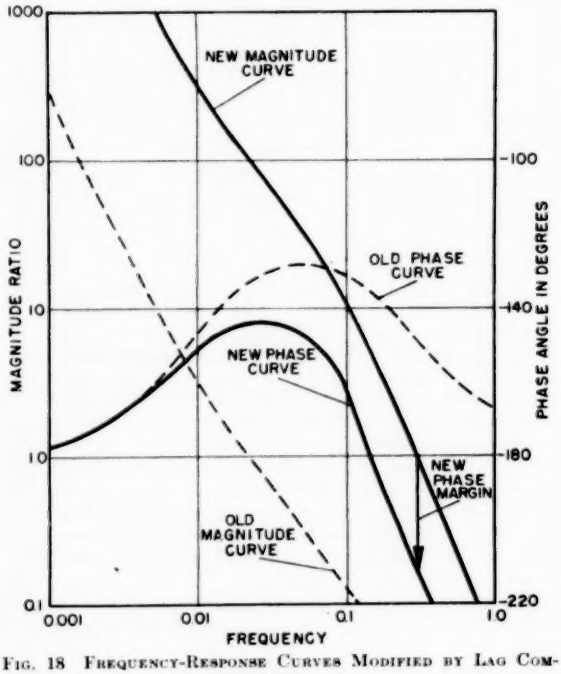


FIG. 18 FREQUENCY-RESPONSE CURVES MODIFIED BY LAG COMPONENT

margin at P is by definition $1/R$ if $R < 1$ and R if $R \geq 1$.

Design Rule 4. If after gain crossover the slope of the magnitude curve is more than —2 (curve not as steep as when the slope is —2) until a point P is reached, and the slope eventually is less than —2 (steeper) the magnitude margin at P should be 2 (6 decibels) or more.

The point P is shown on the dashed magnitude curve in Fig. 11, and the magnitude margin QP at P is 3 (9 decibels). On the solid curve this margin is inadequate.

On an asymptotic curve the point P is uniquely determined and Design Rule 4 is then particularly easy to apply.

We noted previously that Bode's phase-magnitude relationship implies that increasing the slope of the magnitude curve (making it less steep where the curve is falling) increases the phase angle. Thus if the point of gain crossover with which the phase margin is associated is kept fixed and the slope of the magnitude curve is increased (less steep where the curve is falling), *the phase margin is improved.* From the same Bode relationship it follows that if the slope of the magnitude curve never is less than —2 (no steeper on the falling portions), the phase lag never exceeds 180 deg, and there is no gain margin. Any portion of a magnitude curve where the slope is less than —2 (steeper) tends to increase the phase lags and deteriorate the phase margin.

Innumerable design rules like those of this section can be formulated. No group of such rules can replace the direct mathematical study of a given problem, but they can help the engineer in his rough analysis of a control problem.

For higher frequencies lag components tend to decrease phase angles whereas lead components tend to increase them.

The frequency-response curves for an example of a simple lead component are shown in Fig. 15. Adding ordinates, the response curves of the system of Fig. 12(*b*) change to those of Fig. 16 when this lead component is incorporated.

For a lag component we have the corresponding curves in Fig. 17, and the result of adding ordinates to those of Fig. 12(*b*) yields the curves of Fig. 18 for the system obtained. At higher frequencies the new magnitude curve drops below the old one.

The system of Fig. 16 is stable, whereas that of Fig. 18 is unstable.

Systems Without Unity Feedback

Consider the system of Fig. 19 where there is a component in the feedback path so that the input to the control is not $r - c$ as in Fig. 10.

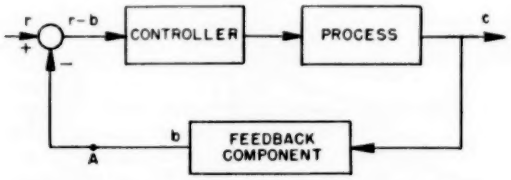


Fig. 19 System With Component in Feedback Path

The previous theory applies here, but to obtain the open-loop magnitude-frequency-response curve, as when we open the loop at A, we must add ordinates of the three magnitude curves for the controller, process, and feedback component. The same is true for the phase curve. The case of Fig. 19 is a common one since the variable c to be controlled must be measured and the measuring element will generally involve leads and lags. The controller responds to the difference $r—b$ between the input r and the output b of the feedback component. Thus a temperature control with a thermocouple measuring element is sensitive to the difference between the temperature setting and the temperature as measured by the thermocouple. The thermocouple has a thermal lag.

If for the range of frequencies involved in the stability of the loop shown in Fig. 19, the response of the feedback component can be neglected, the loop of Fig. 19 reduces to that of Fig. 10. This could be the case for a temperature control with a fast thermocouple.

Transfer Locus (Nyquist) Plots

The DS-Committee does not recommend standardizing on any one type of frequency-response replot. The type of replot that most clearly demonstrates the effects of a component on over-all performance depends on the location and character of the component and upon the criteria of performance desired. The application of frequency response to some common problems is well understood. However, its application to complex control problems, particularly process control, has not been evaluated fully. Industrial processes are subject to multiple disturbances and interacting effects, many of which are known to exist, but few of which have been quantitatively determined. Means of demonstrating these effects will be developed by the practicing engineer. In publishing his findings, he should be free to choose and devise the replots best-suited to his problem. However, since the committee recommends that the magnitude and phase frequency-response curves be given on logarithmic co-ordinates, the reader will find it convenient to use the log magnitude-phase charts to be discussed later. The DS-Committee therefore makes the following suggestion:

Recommendation 8. Where it is convenient to do so the use of log magnitude-phase charts (also referred to as Black or Nichols charts) in preference to other types of replots is recommended.

The *transfer locus plot* is a commonly used and widely recognized replot. This is the well-known Nyquist plot and combines in one curve the information supplied by the magnitude and frequency-response curves of an open loop (the other replots are also called "transfer locus plots" in the literature, but we prefer to reserve this terminology for the Nyquist plots). This plot is obtained by drawing a curve on polar co-ordinates where the vectorial angle for a frequency f is the open-loop phase angle, and the length of the radius vector is the open-loop magnitude ratio for the frequency f. The frequency is thus the parameter along the curve.

The solid curves in Fig. 11 for the governor-engine open loop combine to give the transfer locus of Fig. 20. Thus when $f = 0.064$, the magnitude ratio is 3 and the phase angle —136 deg as shown for the point P. The light portions of the curves in Fig. 20 to the left of the vertical axis are distorted to show how the curves are asymptotic to the vertical axis. Also, the curves are enlarged in the neighborhood of the origin to make the small loop visible.

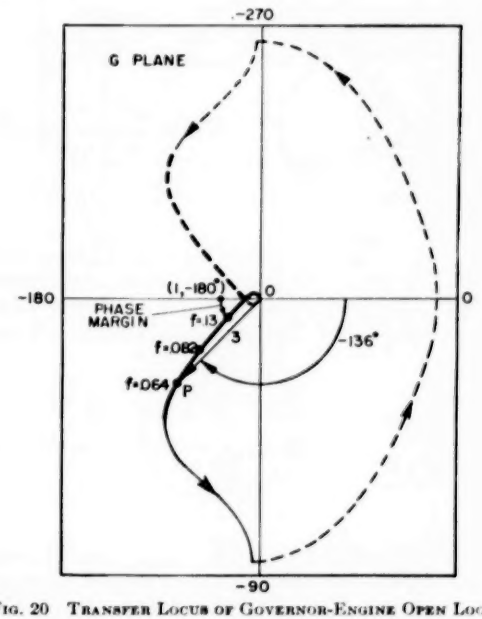


Fig. 20 Transfer Locus of Governor-Engine Open Loop

To enlarge the scale of the pertinent portion of the plane the DS-Committee makes the following recommendation:

Recommendation 4. Where readability and accuracy of data extraction can be improved by moving the co-ordinate axes to one side or corner of the transfer locus (Nyquist) plot, this is recommended.

Mathematical Background

It is hoped that even though the mathematically untrained reader may not understand the details of the present section he will at least obtain a general idea of the theory involved.

The dependency of the output of a linear system on the input (11) is described by the "transfer function" $G(s)$ of a complex variable s, where $G(s)$ is related to the concept of impedance in electrical theory. It can be shown that the transfer locus is a plot of this function for $s = j\omega$, where $j = \sqrt{-1}$, and $\omega = 2\pi f$ for the frequency f, while f varies from 0 to ∞. By drawing a symmetrical branch with respect to the horizontal axis the conjugate curve is obtained, associated with $G(j\omega)$ for ω ranging from 0 to $-\infty$. For the governor-engine example the conjugate curve is shown dashed in Fig. 20 as a branch running from the origin upward and asymptotic to the —270-deg ray. For the transfer functions normally treated (such as those analytic except for poles) theory shows that if the magnitude of the radius vector increases beyond all bounds as ω goes to zero, the transfer locus is asymptotic to one of the vertical or horizontal rays, emanating from the origin O. If the transfer locus terminates at finite points, the locus and its conjugate form a closed curve. The end of the transfer locus associated with $\omega = \infty$ is a finite point (4). If the other end ($\omega = 0$) is asymptotic to one of the rays emanating from the origin, the curve formed by the transfer locus and its conjugate is closed by drawing a properly chosen arc. In physical systems the locus is often asymptotic to the —90-deg ray and sometimes to the —180-deg ray. It is fairly uncommon to have the locus asymptotic to the —270-deg ray. It therefore will be sufficient for what follows to show the reader how to close the curve formed by a transfer locus and its conjugate in the cases where the —90, —180, or —270-deg ray is the asymptote.

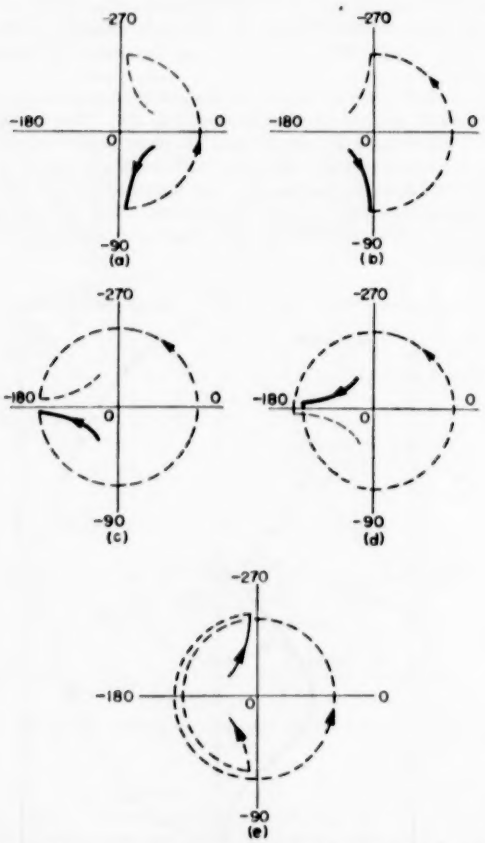


Fig. 21 Methods of Closing a Curve Formed by Transfer Locus and Its Conjugate

From complex variable theory involving the Nyquist stability criterion the curves must be closed in a certain way (the closed curve is to be traversed once as one goes once, say, clockwise, "around" the right half of the s-plane) and the arcs for closing the curves can be taken to be circular arcs (they may be more than 360 deg) of very great ("infinite") radius as indicated in Fig. 21. In Figs. 21 (a) and (b) the —90-deg ray is the asymptote. In Figs. 21 (c) and (d) the asymptote is the —180-deg ray. Finally, Fig. 21(e) shows the method of closure in the case of a —270-deg ray. In Fig. 21(e) the transfer locus is asymptotic to the —270-deg ray in the third quadrant, and it takes $1^1/_2$ revolutions in the clockwise sense to close the curve. The curve formed by the transfer locus, its conjugate, and closure arc will be called the "closed transfer locus."

In Fig. 20 the closure is shown by the dashed arc in the right half plane. The closed transfer locus can be traversed in a counterclockwise direction as shown by the arrows. As we traverse this locus we do not encircle point (1, —180 deg).

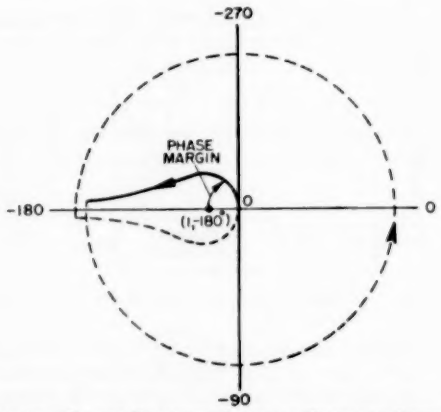


Fig. 22 Closed Transfer Locus of Unstable System

For the transfer locus of Fig. 22 plotted as a solid line the curve is closed as shown. When we traverse the curve once as indicated by the arrows we encircle the point (1, —180 deg) twice. The closed transfer locus for the example of Fig. 12(d) is shown in Fig. 23. The *net* encirclement of the point (1, —180 deg) for this case is zero.

To get the *net encirclement* draw a vector from the point (1, —180 deg) to a point P on the transfer locus. Let P traverse the entire closed curve. The vector will rotate through an angle ($360n$) deg, where n is the number of times the locus encircles the point (1, —180 deg).

It is assumed that the closed-loop linear systems under discussion are stable if opened. For such systems Nyquist's famous stability criterion follows:

Nyquist Stability Criterion. Let S be a closed-loop system that is stable when the loop is opened. The system S is stable if and only

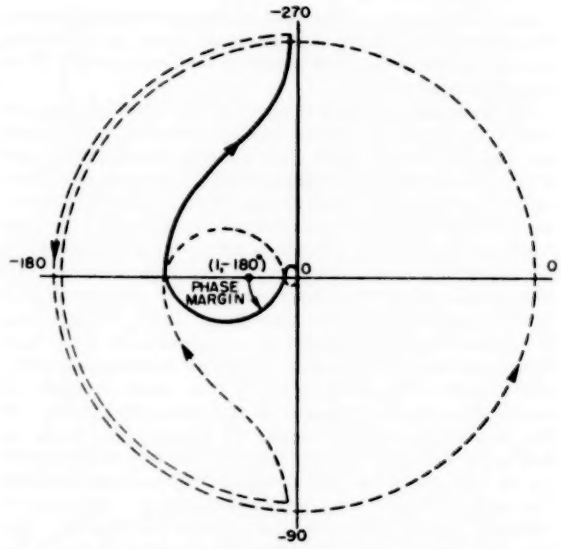


FIG. 23 CLOSED TRANSFER LOCUS FOR CASE OF FIG. 12(*d*)—A STABLE SYSTEM

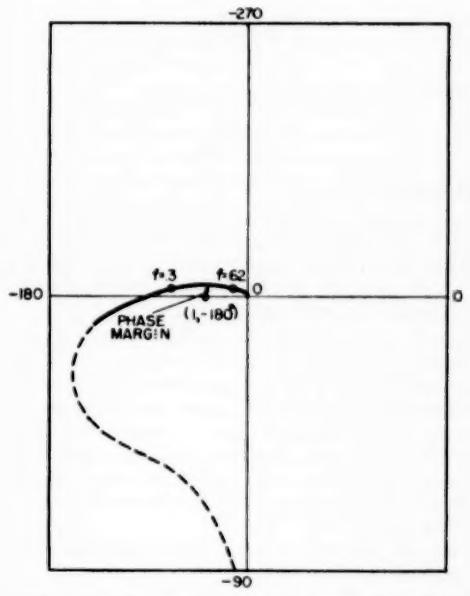


FIG. 24 TRANSFER LOCUS OF UNSTABLE SYSTEM WITH HIGH GAIN

if in one traversal of the closed transfer locus the net encirclement of the point (1, —180 deg) is zero.

The cases of Figs. 20 and 23 are stable by this criterion whereas Fig. 22 is the locus of an unstable system. The system S is unstable if the locus passes through the point (1, —180 deg).

The Nyquist stability criterion can be stated in a more general form to cover systems with unstable open loops.

DESIGN WITH TRANSFER LOCUS PLOTS

For the practicing engineer, Nyquist's criterion can be replaced by the following:

Practical Stability Criterion. A system is stable if it has a positive phase margin. Otherwise it is unstable.

This criterion has been proved to hold for a system S that is "absolutely stable" (3), that is, a system that is stable up to some value of gain and unstable for larger values, or that is stable for all gains.

Multiplying the gain of the open loop by a factor k corresponds merely to multiplying each radius vector by the factor k. Thus raising the gain by the factor 10 in the governor-engine example changes the transfer locus of Fig. 20 to that of Fig. 24. The locus now loops around the point (1, —180 deg), and the closed system is unstable. The dashed portion of the curve in Fig. 24 is a distorted section to show how the curve is asymptotic to the vertical axis. The phase margin is seen to be negative as shown.

M AND N LOCI

With the aid of auxiliary curves the frequency response of the closed loop of Fig. 10 can be obtained directly from the transfer locus of the open loop.

Let M be the magnitude ratio of the closed loop (see Fig. 10 with r as the input and c as the output). It can be shown that the curves of constant M are circles on the plane of the transfer locus, as shown in Fig. 25. Here the transfer locus of Fig. 20 is drawn on the same graph. When $f = 0.083$ the locus crosses the curve $M = 1.5$. Hence the magnitude ratio of output c to input r at $f = 0.083$ is 1.5 for the closed loop. Note that the "maximum value" M_p of M for the points on the transfer locus is 2.23. It can be shown that if the maximum value M_p of M

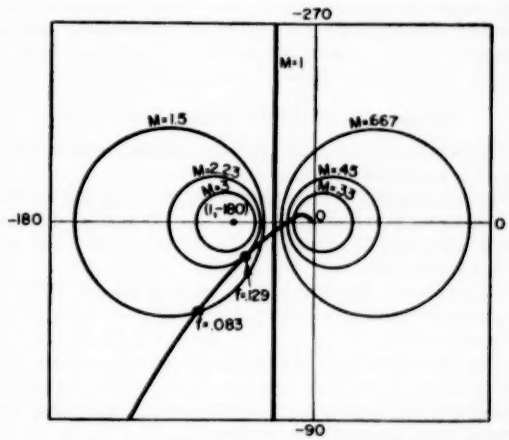


FIG. 25 M-CIRCLES ON PLANE OF A TRANSFER LOCUS

for a transfer locus plot is very large such as 4, the transient performance of the closed loop is poor. A transfer locus near the instability point (1, —180 deg) implies a high M_p.

For $M > 1$ an M circle has the radius

$$\frac{M}{M^2 - 1}$$

and center on the horizontal axis a distance

$$\frac{M^2}{M^2 - 1}$$

to the left of the origin. If $M = 1$ the M-circle reduces to a straight line $\frac{1}{2}$ unit to the left of the vertical axis. If $M < 1$ the M-circle has the radius

$$\frac{M}{1 - M^2}$$

and center on the horizontal axis a distance

$$\frac{M^2}{1 - M^2}$$

to the right of the origin.

The following design rule is in common use (12). If the frequency at which the maximum M is attained is positive, this frequency is termed the resonant frequency. If at zero frequency $M = 1$, the value of the frequency at which M first falls to 0.707 is termed the *band width*.

Design Rule 5. The maximum magnitude ratio M_p of a closed loop should be less than 2 and should be about 1.3. The band width or resonant frequency (if there is one) should be high as possible.

Mathematical considerations show that a low resonant frequency or band width means slow transient performance. The band width (or resonant frequency) in Design Rule 5 is limited by "noise" suppression requirements and other considerations.

Curves of constant phase lag for closed-loop frequency response are circles, and are called "*N-loci.*"

NOTE: By Design Rule 5 the dashed magnitude curve of Fig. 11 is preferred to the solid one.

The quantity M_p is related to the phase margin (see Design Rule 1). Suppose that gain crossover occurs at a point where the phase margin is 30 deg. The value of M at this point is 1.93 (approximately), whence M_p is at least 1.93. As the phase margin is decreased the value of M at gain crossover increases. Thus if M_p is less than 1.9 the phase margin will be greater than 30 deg. Similarly, if $M_p = 1.3$ the phase margin is over 45 deg.

Recommendations for Transfer Locus Plots

The DS-Committee makes the following recommendation with regard to transfer locus plots:

Recommendation 5. It is recommended that a sufficient number of frequencies be shown on the transfer locus to indicate the rate at which frequency affects the locus. The use of M and N-loci should be kept to minimum.

The recommendation to keep the use of M and N-loci to a minimum is made to prevent the drawing from being crowded with too many curves.

It has been customary to write the transfer function associated with the open loop as KG for a gain constant K. It is often inconvenient to determine and remove K from the transfer function, especially in experimental studies. Instead, the following is suggested:

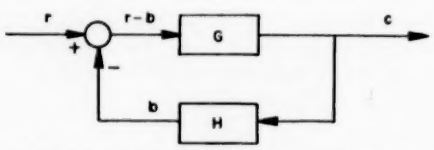


FIG. 26 LOOP WITH BOTH FORWARD AND FEEDBACK COMPONENTS

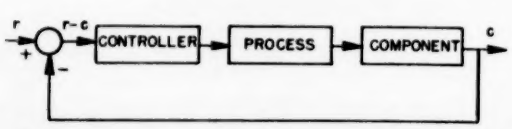


FIG. 27 SYSTEM OF FIG. 19 WITH FEEDBACK COMPONENT PLACED IN FORWARD PORTION

Recommendation 6. Let G denote the open-loop transfer function. The plane of the transfer locus should be labeled as the G-plane (see Fig. 20) rather than as the KG-plane for a gain constant K and frequency variant factor G.

The preceding discussion of transfer locus plots has been restricted to closed-loop systems without frequency variant components in the feedback path. Let G denote the transfer function of the forward portion of the loop, namely, of the controller-process in Fig. 19. Let H designate the corresponding function for the feedback component so that GH is the transfer function of the open loop (see Fig. 26). As noted before, the magnitude curve (on logarithmic co-ordinates) of the open loop is obtained by adding ordinates of the magnitude curves of the forward and feedback portions, and the same thing is true of the phase. In so far as the stability question is concerned, the reader may treat the system as if all components were in the forward portion, as in Fig. 27. The M and N-loci now no longer designate the magnitude and phase response of the closed loop of Fig. 19. The designer can, however, design the elements involved in the problem so that good stability is secured for the arrangement of Fig. 27, using the criterion for maximum M given in the foregoing. The system of Fig. 19 will now be equally stable. However, the frequency response of the two loops will not be identical. This may be a factor, especially where it is necessary to have the output c follow the input r very precisely, as in aiming an anti-aircraft gun.

Thus far we have focused our attention on the response c to input settings r of the controller. This is of prime importance in the design of follow-up servos. In the usual regulator applications, however, a variable is kept at a constant or reasonably constant value and the system is subjected to disturbances not arising from the controller setting. These are often load disturbances, and will be referred to as such. The transfer function relating the load L and the output c will, in general, be different from that relating r and c. However, the stability problem is the same. Thus instead of treating the response of the closed loop to sinusoidal oscillations of the load, we consider the response to sinusoidal oscillations of the controller setting. There are, of course, situations where the designer will wish to study the response to load oscillations directly.

Inverse Nyquist Plots

The transfer locus plot is the plot of the open-loop transfer function G. On the other hand, the inverse Nyquist plot (13) is the polar plot of $1/G$, and is obtained as follows: The magnitude of the radius vector on the inverse Nyquist plot is the magnitude (amplitude) of the input to the open loop divided by the magnitude of the output (the reciprocal is used for the transfer locus plot) and the vectorial angle is the negative of the phase angle used for the transfer locus plot.

For the case of unity feedback shown in Fig. 10, the inverse Nyquist plot for the closed loop (relating r and c) is obtained by a simple displacement of the points of this plot for the open loop. The inverse plot makes it easy to modify the plot for the incorporation of feedback components. These considerations and its emphasis on the low-frequency region often make the inverse plot a convenient one to work with. The inverse Nyquist plot for the governor-engine example of Fig. 20 is shown as the solid curve in Fig. 28. The reader will note that the locus is above the (1, 180 deg) point instead of below it as for the direct Nyquist plot. The phase margin is indicated in the figure. To insure stability it is required that this margin be positive, and therefore that the curve, assuming a phase margin exists, must pass *above* the (1, 180 deg) point as shown. The inverse plot for the closed loop is given by the dashed line in Fig. 28, obtained by displacing each point one unit to the right. For the case of unity feedback

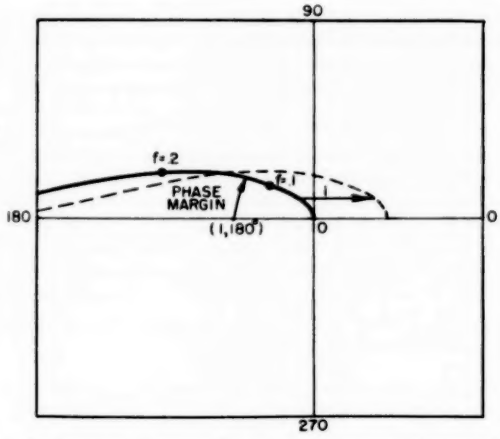


FIG. 28 INVERSE NYQUIST PLOT FOR GOVERNOR-ENGINE CASE

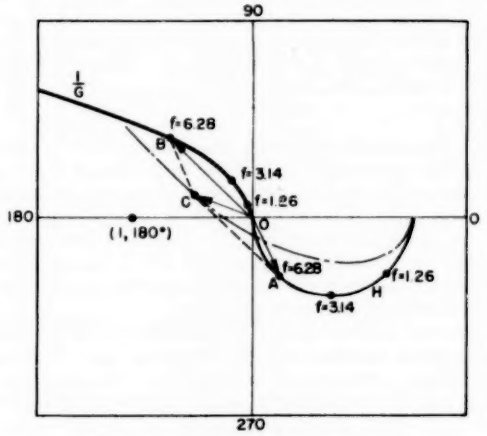


FIG. 29 INVERSE NYQUIST PLOTS

this displacement thus yields the closed-loop plot from the open loop.

In Fig. 29 are shown inverse Nyquist plots for a physical system with G and H as forward and feedback transfer functions (see Fig. 26). The heavier curve marked $1/G$ is the inverse Nyquist plot of the forward part of the loop, and the lighter one marked H is the transfer locus for the feedback portion. The points marked A and B are points for the same frequency f on the H and $1/G$ curves, respectively, where $f = 6.28$. Adding the vectors OB and OA, we obtain the point C on the inverse Nyquist plot of the closed loop by drawing the diagonal OC of the parallelogram with sides OA and OB. The closed-loop inverse locus is the curve with long and short dashes in the figure.

The M-loci are concentric circles on the inverse plot. The distance from the origin to a point on the inverse Nyquist plot for the closed loop is simply $1/M$ for the magnitude ratio M defined as for the direct Nyquist plot (output magnitude over input). The inverse transfer function for the closed loop is simply

$$\frac{1}{G} + H$$

Recommendation 7. Let G and H be the respective transfer functions of the forward and feedback portions of a closed loop. On the inverse Nyquist plot the loci of the forward and feedback parts of the loop should be labeled 1/G and H, respectively (see Fig. 29). The use of M-circles should be minimized.

LOG MAGNITUDE-PHASE PLOTS

The log magnitude-phase (9) plot is obtained by graphing magnitude ratio as the vertical co-ordinate on a logarithmic scale, and phase angle on a horizontal linear scale. The log magnitude-phase plot has the following advantages:

(*a*) The basic frequency-response data, presented with the aid of logarithmic co-ordinates as recommended in this paper, may be transferred directly to the log magnitude-phase plot with dividers.

(*b*) The products of several component frequency responses may be combined by graphical addition and thus the effects of components on over-all performance may be easily demonstrated.

(*c*) The log magnitude-phase plot has the advantage that the gain of the open loop can be changed by merely sliding the log magnitude-phase plot up or down. Changing the gain does not affect the phase angle but moves the plot up or down as described for the magnitude curve on logarithmic co-ordinates. Normally, a co-ordinate system with M and N-curves on a template is slid up and down over the log magnitude-phase plot to change the gain of the loop.

The log magnitude-phase plot has the distinct disadvantage that M-contours are not circles; hence special construction, or templates, are necessary for drawing them. For the governor-engine example of Fig. 11 the log magnitude-phase plot is shown in Fig. 30.

Recommendation 8. It is recommended that the log magnitude-phase plot be graphed with the magnitude ratio as the vertical co-ordinate on a logarithmic scale and phase angle in degrees on a linear horizontal scale.

TRANSFER FUNCTIONS

Suppose that a physical system S is at rest before the instant

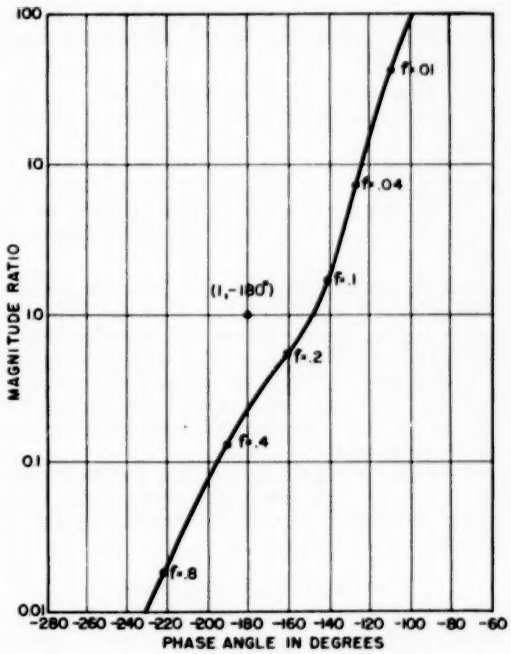


FIG. 30 LOG MAGNITUDE-PHASE PLOT FOR GOVERNOR-ENGINE PROBLEM

$t = 0$. Let the input m to S be one which vanishes identically for $t < 0$. Let c denote the output of S, and assume that the Laplace transforms of m and c exist.

The *transfer function* of S is by definition the Laplace transform of the output divided by that of the input, provided that this quotient is unique.

Recommendation 9. The independent variable in a transfer function should be denoted by s. Thus $G(s)$ is a transfer function of the complex variable s. In actual computation $j\omega$ is to be substituted for s.

Recommendations Regarding Presentation of Measurement Data

It is recommended that from the point of view of the presentation of measurements of frequency responses, the least common denominator of technical audiences should govern.

Since attention should be focused on the diagram of the system studied, it is not desirable to load up this drawing. If the measurement equipment is simple, this equipment may be indicated on the diagram with symbols different from those used to describe the system components, or by the use of dotted lines, indicating at most the real or equivalent circuit of the measurement system. If the measurement equipment is complicated, a simple diagram may be repeated, with the measurement system put in by dotted lines. The characteristics and type of measurement systems used may be tabulated either in the text or in the diagram for the measurement system. Most extensively, *these data could include:*

(*a*) *The input impedance* in mechanical or electrical terms of the measurement system relative to the impedance of the primary system across the points of measurement.

(*b*) *The input-output sensitivity* of the measurement system as a whole or by components.

(*c*) *The equivalent circuit of the measurement system*, its *frequency response*, or its *time constant* if this exists.

The DS-Committee has been concerned with the problem of insuring that the measurement methods used in obtaining frequency-response data:

(*a*) Can be reproduced to obtain the same data.

(*b*) Result in no modifications of the "true" frequency response.

(*c*) If they modify the true response, that it may be corrected for.

(*d*) Or if it cannot be corrected for, that suitable discussion be included to indicate the qualitative effect of the measurement methods.

As much information should be given as would permit the casual reader to have a qualitative idea of the significance of the measuring systems, and the expert a quantitative idea. To achieve this, such statements as "a rotating strut-type wire-fatigue-testing machine," "a linear accelerometer with a wide frequency band," or "the system response has no appreciable lag" are not adequate. However, not much more information is needed to be sufficient.

The common-sense view should prevail so that the description is suited to the principal uses of the system, with particular references to the limits of operation, and the effect of uncontrolled factors in the ambient environment.

Additional Recommendations

The frequency-response data should include the range of inputs for which the data are linear, and experimental or theoretical *estimates of nonlinearity* should be given where possible. In general, if a large sinusoidal input signal and a small input sinusoid produce the same magnitude and phase responses, the process can be treated as linear, at least at the control point about which the measurement was taken. If the linearity is not known, a detailed description of the experimental setup should be given.

Failure of Design Criteria

Commonly used criteria for stability may be satisfied while the system is unstable for practical purposes. To illustrate, consider the system with the open-loop transfer function G given by

$$G = \frac{0.486}{s^2 + 0.3s + 1} \quad [1]$$

The system is taken to have unity feedback. For the function G we have the following:

$$M_p = 1.34 \quad \text{at } \omega_r = 1.2$$

Phase margin: 45 deg

Gain margin: ∞

Here ω_r is 2π times the resonant frequency. However, the roots of the equation

$$1 + G = 0 \quad [2]$$

are

$$-0.15 \pm 1.21j \quad [3]$$

The reader familiar with the root method of treating stability will see that the transients will be highly oscillatory because the coefficient 1.21 of j is much greater than the numerical value 0.15 of the real part. It is the experience at the author's company that if the real part is numerically small, such as 0.1, and the coefficient of j is about 10 times the real part or more, the system is unstable. This occurs in the governing of the speeds of prime movers where 1 sec is the unit of time used.

The present example violates Design Rule 2 since the slope of the magnitude curve on log co-ordinates at gain crossover is approximately —4, as shown in Fig. 12 (*a*). However, *this rule was made only to insure that the phase margin would be adequate*, which it is. Thus *satisfying the stability criteria will not insure adequate stability or adequate transients*, but may help to achieve this.

The Role of Frequency-Response Approach

The frequency-response approach has attained its present position in the automatic-control field partly because many engineers prefer working with curves rather than algebraic formulas, and there is a general lack of familiarity with simple methods (14) for solving algebraic equations of high degree. Where possible it is desirable to derive the actual transients, or at least obtain the characteristic roots. An analyst familiar with the theory of such roots can tell from them what the curves for the transients will look like. Where an analog computer is available the problem of obtaining the transients is an easy one. Considerable use is now being made of the Evans root-locus method (15) of solving an algebraic equation. Like the frequency-response design methods it has the disadvantage that it involves the graphing of curves. All of the curve-drawing approaches to the design of automatic controls, including the analog computer, suffer from the fact that a curve represents only a numerical case. Correct and rapid design can often, if not generally, be accomplished only by manipulating the mathematical formulas with general symbols for some of the constants. To do this a considerable amount of mathematical training is indicated. Actually, both numerical and abstract studies are needed in the analysis of automatic controls.

The great power of the frequency-response approach for the control engineer or scientist, with a considerable mathematical background, lies in the application of frequency-response methods to areas, such as nonlinear phenomena, where the usual algebraic methods fail (16). Hysteresis, saturation, and variable damping are examples. The use of frequency-response curves for diagnostic purposes rather than design is also invaluable in that factors that should not have been neglected in a problem often show up on these curves.

For the beginner the frequency-response approach is a boon. This is especially true because of the correlation that exists between the transient response and the frequency response of a physical system (7).

Appendix

To clarify the use of transfer functions in making design rules we shall consider a system with the open-loop transfer function

$$\frac{k(s + a)}{s^2(s + 1)} \quad [4]$$

where a and k are positive real constants, and $a < 1$. The asymptotic magnitude curve in this case is composed of three falling straight-line sections with slopes -2, -1, and -2, respectively, Fig. 14. The magnitude ratio R for this function is given by

$$R = \frac{k\sqrt{a^2 + \omega^2}}{\omega^2\sqrt{1 + \omega^2}} \quad [5]$$

The ratio p of the value of R at $\omega = a$ to the value of R at $\omega = 1$ is given by

$$p = \frac{2}{a(1 + a^2)} \quad [6]$$

The characteristic equation for the closed loop is

$$x^3 + x^2 + kx + ak = 0 \quad [7]$$

Routh's stability criterion (6) implies that $1 > a$. For a value of a "near" 1, such as 0.5, the roots of the foregoing equation are much too oscillatory. For $a = 0.5$ we have $p = 3.2$. For $a = 0.2$ the roots are satisfactory and $p = 9.6$. For these and other reasons we have recommended in Design Rule 3 that p be chosen equal to or greater than 5.

BIBLIOGRAPHY

1 "Communication Networks," by Ernst A. Guillemin, John Wiley & Sons, Inc., New York, N. Y., vol. 2, 1935, p. 475.

2 "Regeneration Theory," by Harry Nyquist, *Bell System Technical Journal*, vol. 11, 1932, pp. 126–147.

3 "Network Analysis and Feedback Amplifier Design," by Hendrik Bode, D. Van Nostrand Company, Inc., New York, N. Y., 1945.

4 "Mathematical Engineering Analysis," by Rufus Oldenburger, The Macmillan Company, New York, N. Y., 1950.

5 "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948.

6 "Servomechanism Analysis," by G. J. Thaler and R. G. Brown, McGraw-Hill Book Company, Inc., New York, N. Y., 1953.

7 "Servomechanisms and Regulating System Design," by Harold Chestnut and R. W. Mayer, John Wiley & Sons, Inc., New York, N. Y., 1951.

8 "Introduction to Servomechanisms," by A. Porter, Methuen & Company, Ltd., London, England, 1950.

9 "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

10 "The Dynamics of Automatic Controls," by R. C. Oldenbourg and H. Sartorius, The American Society of Mechanical Engineers, New York, N. Y., 1948.

11 "Transients in Linear Systems," by M. F. Gardner and J. L. Barnes, John Wiley & Sons, Inc., New York, N. Y., vol. 1, 1942.

12 "Application of Circuit Theory to the Design of Servomechanisms," by A. C. Hall, *Journal of The Franklin Institute*, vol. 242, October, 1946, p. 287.

13 "Parallel Circuits in Servomechanisms," by H. T. Marcy, Trans. AIEE, vol. 65, 1946, pp. 521–529.

14 "Practical Computational Methods in the Solution of Equations," by Rufus Oldenburger, *American Mathematical Monthly*, vol. 45, June-July, 1948, pp. 335–342.

15 "Control System Synthesis by Root Locus Method," by W. R. Evans, Trans. AIEE, vol. 69, part I, 1950, pp. 66–69.

16 "Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements," by E. Calvin Johnson, Trans. AIEE, vol. 71, part II, July, 1952, pp. 169–181.

Summary of Committee Recommendations

1 It is recommended that magnitude curves be plotted on logarithmic co-ordinates with the magnitude ratio as the vertical co-ordinate and the frequency in cycles per unit time as the horizontal co-ordinate.

2 It is recommended that phase be plotted in degrees on a linear vertical scale and frequency in cycles per unit time on a horizontal logarithmic scale.

3 Where it is convenient to do so the use of log magnitude-phase charts (also referred to as Black or Nichols charts) in preference to other types of replots is recommended.

4 Where readability and accuracy of data extraction can be improved by moving the co-ordinate axes to one side or corner of the transfer locus (Nyquist) plot, this is recommended.

5 It is recommended that a sufficient number of frequencies be shown on the transfer locus to indicate the rate at which frequency affects the locus. The use of M and N-loci should be kept to a minimum.

6 Let G denote the open-loop transfer function. The plane of the transfer locus should be labeled as the G-plane (see Fig. 20), rather than as the KG-plane for a gain constant K and frequency variant factor G.

7 Let G and H be the respective transfer functions of the forward and feedback portions of a closed loop. On the inverse Nyquist plot the loci of the forward and feedback parts of the loop should be labeled $1/G$ and H, respectively (see Fig. 29). The use of M-circles should be minimized.

8 It is recommended that the log magnitude-phase plot be graphed with the magnitude ratio as the vertical co-ordinate on a logarithmic scale and phase angle in degrees on a linear horizontal scale.

9 The independent variable in a transfer function should be denoted by s. Thus $G(s)$ is a transfer function of the complex variable s. In actual computation $j\omega$ is to be substituted for s.

ASME-IRD Dynamic Systems Committee

R. Oldenburger—Chairman
S. P. Higgins—Secretary

W. R. Ahrendt	A. S. Iberall
P. Buckley	M. V. Long
W. I. Caldwell	R. S. Lovett
G. H. Cohen	C. E. Mason
P. S. Creager	C. D Pengelley
G. F. Gardner	D. W. St. Clair
J. A. Hrones	W. E. Vannah
R. M. Hutchinson	E. H. Woodhull

F. A. Woods

Summary of Design and Stability Criteria

Principal Design Criteria

Design Rule 1. The phase margin should be at least 30 deg and the gain margin at least 2.5 (8 decibels).

Design Rule 2. The slope of the magnitude curve on logarithmic co-ordinates at and near gain crossover should be about —1

Design Rule 5. The maximum magnitude ratio M_p of a closed loop should be less than 2, and should be about 1.3. The band width or resonant frequency (if there is one) should be high as possible.

Some Other Design Criteria

Design Rule 3. Consider a system with a portion of its open-loop magnitude curve approximated by the curve of Fig. 14 (composed of straight-line sections), where the sections with slope —2 are long compared to the middle section (section with gain crossover) with slope —1. For this system it is desirable to have the distance AC equal to about $1/3$ or $1/2$ of the distance AB. The quotient of the magnitude ratio associated with the point D by that for E should be at least 5 (the ratio of the frequency associated with B to that for A should be at least 5).

Design Rule 4. If after gain crossover the slope of the magnitude curve is more than —2 (curve not as steep as when the slope is —2) until a point P is reached, and the slope eventually is less than —2 (steeper) the magnitude margin at P should be 2 (6 decibels) or more.

Stability Criteria

Nyquist Stability Criterion. Let S be a closed-loop system that is stable when the loop is opened. The system S is stable if and only if in one traversal of the closed transfer locus the net encirclement of the point (1, —180 deg) is zero.

Practical Stability Criterion. A system is stable if it has a positive phase margin. Otherwise it is unstable.

Discussion

P. NASLIN.[7] The author recommends that the M and N-curves be drawn on a template that may be slid up and down over the log magnitude-phase plot. We personally prefer to plot the attenuation-phase curve on a piece of tracing paper laid over the Black chart. This method makes it possible to write anything of interest on the tracing paper, such as a particular constant-gain contour, after the gain has been adjusted. It also makes it possible to plot new points if necessary without removing the template.

Apart from the foregoing considerations we believe that the most satisfactory graphical design procedure is based on the combined use of the log magnitude-frequency and phase-frequency curves and of the log magnitude-phase plot, and that there is no need whatsoever for any sort of polar plot except in the form of rough sketches for the investigation of the stability of an unknown system.

A. TUSTIN.[8] There seems to be little that is of a controversial nature in the main recommendations concerning the form in which frequency-response data should be presented. This is especially so on taking account of the general recommendation, with which the writer agrees, that "the practicing engineer should be free to choose and devise the form of graphical presentation best suited to his problem."

However, the author has included in his paper not only the recommendations of the Dynamic Systems Committee on the form of presentation of data, but other material, particularly comments on the experience of himself and his company, on desirable phase and gain margins and acceptable design criteria in relation to the log-gain log-frequency diagram. The writer does not question the value of these comments in themselves, but does suggest that they should be kept separate from ASME recommendations about the form of presentation. They relate to design procedure and design criteria and the discussion of this topic is very incomplete.

The recommendations about forms of presentation seem to include only two points of likely controversy: (a) The elimination of the decibel in favor of the common logarithm of the amplitude ratio, and (b) the measure of frequency in cycles per second f, instead of in radians per second ω.

The writer personally shares the committee's view on decibels and has never seen any usefulness in this context in introducing the arbitrary multiplier of 20. The factor 2 in this multiplier originally had some reference to ratios of powers, which are proportional to squares of quantities, but there is no such reference in this context which involves ratios of simple amplitudes, the square root of which has no significance. It is easier and more correct to say "a log-gain of two" than to say "a gain of forty decibels." The writer does not think that it is too late to make this change from the practice of using decibels which is now more prevalent.

On the issue, however, of the unit of frequency, the writer does not think the recommendation should be accepted. Both analytical results as universally presented and the numerical calculations based on them are necessarily in terms of radians per unit time, and it is believed that this unit should be used consistently throughout.

The writer recommends using plain English wherever possible instead of technicians' jargon. Much of the difficulty of communication in this field between engineers of different kinds and between them and potential users has come about because control engineers surely have used so much unnecessary technical jargon. As an example of what is meant, recommendations 1, 2, and 3 of the paper can be stated more clearly in some such words as the following:

1, 2 "It is recommended that in presenting graphically the relationship of output to input for a sinusoidally varying input the presentation should be in the form of two graphs, namely, a graph of the logarithm to base 10 of the ratio of amplitude of output to amplitude of input, and a graph of phase angle in degrees taking lead as positive, both as functions of the logarithm to base 10 of the frequency in cycles per second."

3 "A graph may also be used of the common logarithm of the ratio of amplitude of output to amplitude of input as a function of phase in degrees, with values of the frequency marked at suitable points along the curve."

There is certainly a need for some such recommendations as those that the committee has made. One notes, for example, that our friends from ICI, in paper 54—A-22 of the present series, still insist on being odd man out, by plotting everything inside out and upside down, to the general confusion. It is to be hoped that the present discussion will bring differences in practice of this kind to an end.

HAROLD CHESTNUT.[9] The desire of the Dynamic Systems Committee to extend to more mechanical engineers the understanding of the frequency-response methods used in control-system design is a laudable one. Some of the recommendations of the committee will be well received in many quarters in that many of the same symbols are suggested as have been recommended by the AIEE Committee on Feedback Control System Symbols; for example, $G(s)$ and $H(s)$ are common. In this way a greater measure of standardization throughout engineering can be obtained by the use of these symbols.

The recommendations regarding the use of a logarithmic scale for plotting magnitude ratio while using a linear scale for plotting phase angle will make it difficult to present on the same figure the interrelated quantities of magnitude and phase. As an alternative ordinate scale, use of the logarithm of the gain ratio plotted to a linear scale, as is done in current servodesign practice, permits the common use of a linear ordinate scale for both magnitude and phase angle and is suggested for consideration by the committee. Regarding the recommendation that the logarithmic-frequency scale be expressed in cycles per unit of time, it is suggested that one be allowed the option of using radians per unit of time as an alternative.

The design rules outlined seem to be generally satisfactory, although the possibility exists for ambiguity in the definition of gain margin because the same value of gain margin can be used to indicate a magnitude ratio of greater or less than 1.0. The design rules chosen appear to be of the proper order of magnitude. However, they should serve principally as a guide and should not be considered a primary standard of themselves.

The writer is happy that this subject of frequency-response-data representation and methods of analysis has been brought to the attention of the members of the Society. It is valuable to have this information incorporated into the literature and thinking of the Society.

The writer shares with the author the belief that frequency-response methods aid in estimating the effects of nonlinearities and as such are valuable. They form a useful tool which, with experimental and analog computer work, enable the designer to build improved control systems.

[7] Engineer, Laboratoire Central de l'Armement, Paris, France.

[8] Professor, University of Birmingham, Birmingham, England.

[9] Assistant Engineer, General Electric Company, Schenectady, New York.

ABRAHAM FUCHS [10] The author is to be complimented on his fine paper. As with any set of design rules, there usually are differences of opinion concerning their validity.

Design Rule 5 which states in part, "the band width should be as high as possible," might better have read, "the band width should be no wider than necessary to give a system which meets its speed-of-response specification." This change in emphasis from maximizing band width to minimizing band width is based on the following:

(*a*) The wider the band width of the system, the more complex its structure and the more costly its components and construction.

(*b*) The wider the band width of the system, the more likely is it to follow noise signals. This in turn causes a higher average power consumption without necessarily improving the dynamic accuracy of the system.

(*c*) The wider band width system is more likely to interact seriously with structural resonances and other second-order effects which are not readily analyzable.

This interaction can cause serious malfunctioning.

It is true that there are no completely adequate guides to correlating the band width of a system to a particular performance specification, but there is much to be said for designing for the minimum band width compatible with the desired performance specification.

A. LEONHARD.[11] The Nyquist-Bode criterion, as it has been used until now, permits one to determine relatively quickly whether a system is stable or unstable. It also supplies valuable hints concerning changes in system parameters which might be necessary in order to achieve stability. However, the criterion does not tell anything about the degree of stability, for example, about the relative damping of the roots $|\text{Real part/Imaginary part}| = \rho$. The only thing we do know is that, from experience, the phase margin and gain margin should have certain minimum values to insure the desired degree of stability. (According to Design Rule 1 of the paper, the phase margin should be at least 30 deg and the gain margin at least 2.5.) However, the Nyquist-Bode criterion can be expanded very simply so that the degree of stability also can be determined.

The transfer locus (Nyquist plot) of the open loop represents the ratio of output to input, considering both magnitude and phase, for undamped sinusoidal oscillations with frequencies from 0 to ∞. If this locus, at any frequency, goes through the point $(-1 + j0)$ (to be designated from now on by P_0), then the closing of the loop will produce an undamped oscillation of equal frequency, thus causing instability. The condition for stability is that the point P_0 lies to the left of the locus (or, as shown by Nyquist, that it lies outside the locus).

If we now use damped instead of undamped oscillations for our input, then the output also will consist of damped oscillations, having the same frequency and the same degree of damping. In this case it is also possible to calculate the transfer locus, and, in doing this, it is best to assume the same degree of relative damping for all frequencies. We take the transfer function $G = A(j\omega)/B(j\omega)$ and substitute $\omega(j - \rho)$ for $j\omega$. If this curve goes through the point P_0 at a given frequency ω_a, then this means that a damped oscillation with frequency ω_a and relative damping ρ will occur when the loop is closed. The relative damping of the oscillation will be greater or smaller than ρ if P_0 is to the left or to the right of the curve, respectively.

[10] Senior Engineer, Air Arm Division, Westinghouse Electric Corporation, Baltimore, Md.

[11] Visiting Professor, Istanbul Teknik Universite, Istanbul, Turkey. Translated by D. W. Pessen, Minneapolis-Honeywell Regulator Company, Philadelphia, Pa.

In contrast to the conventional method, in which one determined the stability, and where one never quite knew how far one had to remain from P_0, the curve can now go through P_0 if desired.

As will be shown now, when $\omega(j - \rho)$ is used instead of $j\omega$, the magnitude ratio of the output to the input can still be approximated by asymptotic straight-line rays on logarithmic coordinates. In the example to follow, these asymptotes of the curve for various common control elements are compared for $\rho = 0$ and $\rho = 0.414$. The case of $\rho = 0.414 = \tan^{-1} \pi/8$ represents a degree of damping which is sufficient in most cases.

For each case, the formula for the calculation of the phase angle also is given. The calculation of the phase angle entails somewhat more work for $\rho \neq 0$ than for $\rho = 0$. The curve for the magnitude ratio, however, can be found just as quickly whether $\rho = 0$ or not. In fact, the magnitude-ratio curve for $\rho = 0.414$ differs only slightly from that for $\rho = 0$. As a first approximation, the curve calculated or drawn for $\rho = 0$ can even be used, so that only the phase-angle curvene ed be added. Fig. 31 of this

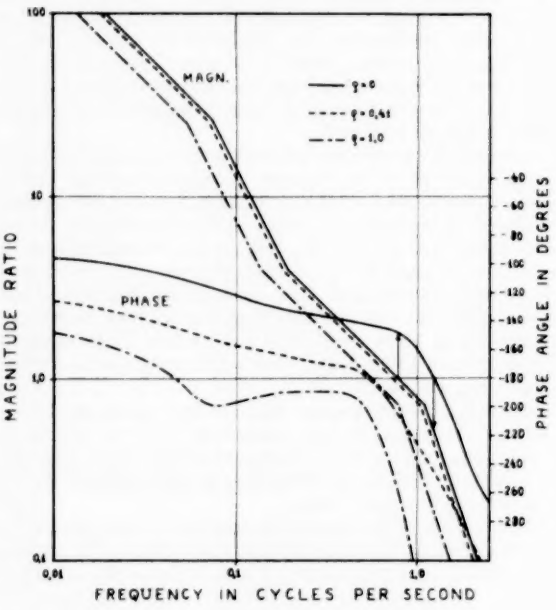


FIG. 31

discussion shows curves for an example in which the open-loop transfer function G is

$$G = K \frac{T_1 s + 1}{T_1 s} \frac{1}{T_2^2 s^2 + 2\zeta T_2 s + 1} \frac{1}{T_3 s + 1}$$

This transfer function might, for example, represent a room-temperature control system, and the first term would represent the proportional-plus-integral-action controller, the second term a steam-water heater with pipe line, and the third term the heated space. For an example such as this, though, the cycles per second in the figure should be changed to cycles per hour. The parameter K is the controller gain, T_1, T_2, and T_3 are time constants, and ζ is a measure of the damping of the second term alone. In calculating the curves, the following values have been assumed: $K = 10$, $T_1 = 0.8$ sec, $T_2 = 0.14$ sec, $T_3 = 2$ sec, $2\zeta = 1.77$.

The curves for $\rho = 0$ (ϕ and M) are those conventionally used to determine stability. One can immediately see that the system is

TABLE 1 TABULATION OF IMPORTANT TRANSFER FUNCTIONS

$\rho = 0$	$\rho = 0.414$
$G_1 = \dfrac{1}{1 + j\omega T}$	$F_1 = \dfrac{1}{1 + \omega(j - \rho)T}$
$\omega_0 = \dfrac{1}{T}$	$\omega_0' = \dfrac{1}{T\sqrt{1 + \rho^2}} = \dfrac{\omega_0}{\sqrt{1 + \rho^2}} = \dfrac{\omega_0}{1.08}$
$\omega < \omega_0$: $\lvert G_1 \rvert = 1.0$	$\omega < \omega_0'$: $\lvert F_1 \rvert = 1.0$
$\left(\omega = \omega_0\text{: } \lvert G_1 \rvert = \dfrac{1}{\sqrt{2}}\right)$	$\left(\omega = \omega_0'\text{: } \lvert F_1 \rvert = \dfrac{1}{\sqrt{2 - \dfrac{2\rho}{\sqrt{1 + \rho^2}}}} = \dfrac{1}{\sqrt{1.24}} = 0.9\right)$
$\omega > \omega_0$: $\lvert G_1 \rvert = \dfrac{1}{\dfrac{\omega}{\omega_0}}$	$\omega > \omega_0'$: $\lvert F_1 \rvert = \dfrac{1}{\dfrac{\omega}{\omega_0'}}$
$\phi = -\tan^{-1}\dfrac{\omega}{\omega_0}$	$\phi = -\tan^{-1}\dfrac{\dfrac{\omega}{\omega_0'}}{\sqrt{1 + \rho^2} - \dfrac{\omega}{\omega_0'}\rho}$
ϕ: $0 \rightarrow -45° \rightarrow -90°$	ϕ: $0 \rightarrow -56° \rightarrow 112.5°$
$G_2 = 1 + j\omega T$	$F_2 = 1 + \omega(j - \rho)T$
$\omega_0 = \dfrac{1}{T}$	$\omega_0' = \dfrac{1}{T\sqrt{1 + \rho^2}} = \dfrac{\omega_0}{\sqrt{1 + \rho^2}} = \dfrac{\omega_0}{1.08}$
$\omega < \omega_0$: $\lvert G_2 \rvert = 1.0$	$\omega < \omega_0'$: $\lvert F_2 \rvert = 1.0$
$(\omega = \omega_0\text{: } \lvert G_2 \rvert = \sqrt{2})$	$\left(\omega = \omega_0'\text{: } \lvert F_2 \rvert = \sqrt{2 - \dfrac{2\rho}{\sqrt{1 + \rho^2}}} = \sqrt{1.24} = 1.11\right)$
$\omega > \omega_0$: $\lvert G_2 \rvert = \dfrac{\omega}{\omega_0}$	$\omega > \omega_0'$: $\lvert F_2 \rvert = \dfrac{\omega}{\omega_0'}$
$\phi = +\tan^{-1}\dfrac{\omega}{\omega_0}$	$\phi = +\tan^{-1}\dfrac{\dfrac{\omega}{\omega_0'}}{\sqrt{1 + \rho^2} - \dfrac{\omega}{\omega_0'}\rho}$
ϕ: $0 \rightarrow +45° \rightarrow +90°$	ϕ: $0 \rightarrow +56° \rightarrow 112.5°$
$G_3 = \dfrac{1}{j\omega T}$	$F_3 = \dfrac{1}{\omega(j - \rho)T}$
$\omega_0 = \dfrac{1}{T}$	$\omega_0' = \dfrac{1}{T\sqrt{1 + \rho^2}} = \dfrac{\omega_0}{\sqrt{1 + \rho^2}} = \dfrac{\omega_0}{1.08}$
$\lvert G_3 \rvert = \dfrac{1}{\dfrac{\omega}{\omega_0}}$	$\lvert F_3 \rvert = \dfrac{1}{\dfrac{\omega}{\omega_0'}}$
$\phi = -90°$	$\phi = -(90° + \tan^{-1}\rho) = -112.5°$

TABLE 1 (*Continued*)

$G_4 = j\omega T$	$F_4 = \omega(j - \rho)T$
$\omega_0 = \dfrac{1}{T}$	$\omega'_0 = \dfrac{1}{T\sqrt{1+\rho^2}} = \dfrac{\omega_0}{\sqrt{1+\rho^2}} = \dfrac{\omega_0}{1.08}$
$\lvert G_4\rvert = \dfrac{\omega}{\omega_0}$	$\lvert F_4\rvert = \dfrac{\omega}{\omega'_0}$
$\phi = +90°$	$\phi = +(90° + \tan^{-1}\rho) = +112.5°$
$G_5 = \dfrac{1}{(j\omega)^2T^2 + j\omega 2\zeta T + 1}$	$F_5 = \dfrac{1}{\omega^2(j-\rho)^2T^2 + \omega(j-\rho)2\zeta T + 1}$
$\omega_0 = \dfrac{1}{T}$	$\omega'_0 = \dfrac{1}{T\sqrt{1+\rho^2}} = \dfrac{\omega_0}{\sqrt{1+\rho^2}} = \dfrac{\omega_0}{1.08}$
$\omega < \omega_0$: $\lvert G_5\rvert = 1.0$	$\omega < \omega'_0$: $\lvert F_5\rvert = 1.0$
$\left(\omega = \omega_0\text{: } \lvert G_5\rvert = \dfrac{1}{2\zeta}\right)$	$\left(\omega = \omega'_0\text{: } \lvert F_5\rvert = \dfrac{1}{2\zeta - \dfrac{2\rho}{\sqrt{1+\rho^2}}} = \dfrac{1}{2\zeta - 0.77}\right)$
$\omega > \omega_0$: $\lvert G_5\rvert = \dfrac{1}{\left(\dfrac{\omega}{\omega_0}\right)^2}$	$\omega > \omega'_0$: $\lvert F_5\rvert = \dfrac{1}{\left(\dfrac{\omega}{\omega'_0}\right)^2}$
$\phi = -\tan^{-1}\dfrac{2\zeta\dfrac{\omega}{\omega_0}}{1 - \left(\dfrac{\omega}{\omega_0}\right)^2}$	$\phi = -\tan^{-1}\dfrac{\dfrac{\omega}{\omega'_0}2\zeta\sqrt{1+\rho^2} - \left(\dfrac{\omega}{\omega'_0}\right)^2 2\rho}{(1+\rho^2) - \dfrac{\omega}{\omega'_0}2\zeta\rho\sqrt{1+\rho^2} - \left(\dfrac{\omega}{\omega_0'}\right)^2(1-\rho^2)}$
ϕ: $0 \rightarrow -90° \rightarrow -180°$	ϕ: $0 \rightarrow -112.5° \rightarrow -225°$

stable, since the phase shift has not yet reached —180 deg at the frequency where the magnitude ratio is 1.

The requirements of Design Rule 1, however, are not quite fulfilled. While the phase margin is 30 deg, the gain margin is only 2.0, as compared to the required value of 2.5. From this one can draw the conclusion that the degree of stability will not be too favorable.

From the curves for $\rho = 0.414$ and $\rho = 1.0$, it can be seen immediately that the relative damping corresponding to those two respective factors is not reached. For both cases, the phase margin is less than 0 ($\phi < -180$ deg). Assuming that all control parameters except the gain K are fixed, then, by reducing K a suitable amount, any desired value of relative damping can be reached. If, for example, we make $K = 7$ instead of 10, then the phase angle remains the same, but the magnitude-ratio curve for $\rho = 0.414$ is displaced downward such that it just goes through 1.0 at the frequency where $\phi = -180$ deg. The phase margin then is zero, which means that a damped oscillation takes place with a relative damping of $\rho = 0.414$ (Real part/Imaginary part of the two predominant roots). Correspondingly, $K = 0.5$ results in a relative damping of $\rho = 1.0$.

The calculation of the asymptotes for various common control elements, and for both $\rho = 0$ and $\rho = 0.414$, is shown in Table 1 of this discussion.

P. S. Creager.[12] The writer wishes to comment, particularly, on recommendations 1 and 2 of the paper, which may be paraphrased as (*a*) plot magnitude ratio versus frequency f on log-log paper, and (*b*) plot "phase" versus frequency f on semilog paper.

The writer has discussed these recommendations with a considerable number of engineers from chemical or petroleum industries. He found a unanimous preference for plotting ω in place of f on a log scale as abscissa. About 75 per cent of these people preferred plotting both "magnitude ratio" and phase on the same sheet. Over 80 per cent preferred "semilog" paper when plotting magnitude ratio alone. They actually plot log of magnitude ratio but have freedom to choose a scale factor and the ordinate of zero value (log = 0) which allows them to portray a wider range of magnitude ratios if desirable.

The net result appears to be that this particular group would prefer to have the recommendation stated somewhat as follows: "Plot 'log of magnitude ratio' and 'phase' on semilog paper with 'ω per unit time' on the logarithmic scale. Plot on same sheet or separate sheets as may be desirable."

The writer would suspect that the subcommittee was unfortunate in not obtaining a more representative sample of opinion in its necessarily limited list of inquiries. He, therefore, hopes that many more opinions may be obtained before recommendations become accepted standards.

Mark Leum.[13] Extreme care always should be taken when recommending standards to avoid any notation or symbolism

[12] Associate Professor of Electrical Engineering, Rutgers University, New Brunswick, N. J.

[13] Mathematician, Woodward Governor Company, Rockford, Ill.

which is not common and could lead to confusion. The system of Fig. 1 of the paper has a transfer function $F(s)$ relating the systems output c and input m. This transfer function $F(s)$ for m sinusoidal may be written $A(\omega)e^{i\varphi(\omega)}$ where ω is the frequency of the input oscillation multiplied by 2π, and $i = \sqrt{-1}$. The quantity $\varphi(\omega)$ is the angular displacement of the input signal m and $A(\omega)$ the absolute value of $F(s)$. Throughout the paper $\varphi(\omega)$ is referred to as simply the "phase" of the transfer function. In just as simple a manner $A(\omega)$ could be referred to as the "magnitude" instead of magnitude ratio.

The vector on the G-plane, Fig. 20, represents the transfer function at a particular frequency. Elementary texts on the subject of vectors usually refer to the length of a vector as the magnitude of the vector. However, this paper recommends as a standard term for the length of a vector magnitude ratio, which appears to be an unnecessary complication.

Another possibility is to replace magnitude ratio by an exact expression for the ratio of output amplitude to input amplitude, that of "amplitude ratio."

Author's Closure

Major Naslin suggests that the magnitude-phase curve be drawn on tracing paper and laid over the Black chart. This recommendation is an excellent one and should save time for the designer. His statement that polar plots generally can be avoided is correct. There is still a school of early workers in the field who were trained to use polar plots and prefer them. This is no doubt in part due to habit and in part to the fact that they already have a library of experimental frequency-response curves expressed as polar plots.

The author agrees with Professor Tustin that the design criteria and committee recommendations are two quite separate aspects of the paper. The author hopes that the inclusion of design criteria will increase the use of committee standards as well as frequency-response methods. The unified approach of the paper should clarify the reasons for the standards, especially to the reader who is not an expert on the subject. Preparatory to writing the paper the author went over the entire available literature on frequency response and formulated Design Rules 1, 2, and 5 on the basis of this study. Design Rules 3 and 4 were employed by the author in his work. Rules 3 and 4 are meant to help the reader in special situations and to indicate that rules can be multiplied ad infinitum. Rules serve as a rough guide for the expert. They are often all that the novice has to work with. Rules can never replace theory. Correct design of automatic controls depends on the entire theory of such controls.

At the end of the paper the committee recommendations are summarized separately from design and stability criteria. For the sake of brevity automatic-control jargon was used in these summaries. Whether or not to avoid technical jargon is always a moot question. Such avoidance tends to make the results available to a larger audience. On the other hand, the necessarily increased length of the statements (we prefer to keep recommendations 1 and 2 separate) works the other way.

After careful study of the subject the DS-Committee decided that the standards should be kept to a minimum. The standards alone did not make sufficient material for an ASME paper. The author therefore hopes that combining the paper with one planned on design criteria was a fortunate decision.

For computations and theoretical development the use of radians per unit time is indicated. However, frequency is understood by a wider class of people. Sine-wave generators and other instruments normally are calibrated in cycles per unit time, rather than in radians. For these and other reasons the DS-Committee prefers frequency.

Mr. Chestnut recommends plotting the logarithm of the magnitude ratio R on a linear scale. It is desirable to list R rather than $\log_{10} R$ as the vertical co-ordinate on the magnitude ratio-frequency plot. Although the values 0.01, 0.1, 1, 10, and so on, of R are equally spaced as shown in Fig. 11 of the paper, this is not true of the intermediate values 2, 3, 4, . . ., 9. The intermediate R markings for $\log_{10} R$ on a linear scale are not such convenient numbers. However, it is not desirable to have two kinds of vertical scales on the same figure and the author concedes that Mr. Chestnut has a point if the magnitude and phase curves are to be plotted as two curves on the same graph. On the other hand, the committee recommends going directly from the magnitude and phase charts to the Black chart where the magnitude and phase curves are combined into one.

Mr. Fuchs is right in saying that the band widths should not be too high. Band width should neither be too low nor too high for the reasons given in the paper.

Professor Leonhard's use of damped oscillations for transfer locus plots is an ingenious departure from the conventional treatment. If the transfer function for the open loop is given it is often not too difficult a matter to replace s by $\omega(j - \rho)$ and plot the corresponding loci for different values of ρ. In this manner the "oscillatory" roots of the characteristic equation may be found. However, the author finds that the methods given in his discussion of the Evans paper,[14] and treated in reference (14) of the present paper, are less time-consuming. Leonhard's plots are valuable for the man who prefers curves to formulas, and should be of considerable help to the student of the subject in giving him a visual picture of the part that the Nyquist plot plays as a special case of Leonhard plots.

Just as the complex representation $e^{j\omega t}$ is employed in place of a real sinusoid $\sin \omega t$ in impedance theory, so we can show that we may use the complex representation $e^{\omega(j-\rho)t}$ in place of the real quality $e^{-\omega\rho t} \sin \omega t$ for a damped oscillation. The quantity $\sin \omega t$ is the imaginary part of $e^{j\omega t}$. Let $E \sin \omega t$ be the forcing function for a linear ordinary differential equation with constant coefficients. The solution (steady state) is then the imaginary part of the solution for the case where the forcing function is replaced by the complex quantity $Ee^{j\omega t}$. That the same thing holds for damped oscillations can be shown by an argument analogous to that given by the author.[15]

Let the given differential equation be

$$O_1(D)v = O_2(D)Ee^{-\omega\rho t} \sin \omega t \quad \ldots\ldots\ldots\ldots [8]$$

where $O_1(D)$ and $O_2(D)$ are polynomials in the derivative D with respect to time. Differentiating both sides of Equation [8] we have

$$O_1(D)v' = O_2(D)\omega Ee^{-\omega\rho t}(\cos \omega t - \rho \sin \omega t) \ldots\ldots [9]$$

Multiplying both sides of Equation [8] by $\omega(j + \rho)$, adding to the corresponding sides of Equation [9], and finally dividing by ω we have

$$O_1(D)\left[\frac{v'}{\omega} + (j + \rho)v\right] = O_2(D)Ee^{\omega(j-\rho)t} \ldots\ldots [10]$$

Clearly, the solution v of Equation [8] is the imaginary part of the solution

$$\left(\frac{v'}{\omega} + \rho v\right) + jv$$

[14] "The Use of Zeros and Poles for Frequency Response or Transient Response," by W. R. Evans, published in this issue, pp. 1335–1344.

[15] Author's bibliography, reference (4), pp. 128–129.

of Equation [10], whereas at the same time $Ee^{-\omega\rho t} \sin \omega t$ is the imaginary part of $Ee^{\omega(j-\rho)t}$.

Although the Leonhard approach has considerable value in so far as theory is concerned, this is not so apparent with regard to the purely experimental aspects of the problem. It is not too difficult a matter to produce sinusoidal forcing functions and obtain the Nyquist plots physically. It does not appear to the author that it is a simple matter to produce forcing damped oscillations of the form $e^{-\omega\rho t} \sin \omega t$ so as to obtain the Leonhard plots experimentally. However, this is certainly worth engineering investigation because the Leonhard plots give much more information than the Nyquist plots.

Professor Creager's comments on the preferences of chemical and petroleum engineers is illuminating. The author hopes that the standards for the presentation of frequency-response data will be reviewed at periodic intervals, such as every five years. Five years should give industry time enough to make a fair test of the standards.

The point brought up by the writer's colleague, Mr. Leum, regarding "magnitude" is an important one. The writer personally prefers the word "amplitude," but the DS-Committee voted otherwise. Technicians are probably generally familiar with the amplitude of a sine wave since this term is already in extensive use. Mr. William Ahrendt correctly points out that amplitude is employed in complex variable theory for the angle or argument of a complex number so that the problem of a confusion of terms exists. Let a linear system be subjected to sinusoidal oscillations. For such a system a transfer function $G(j\omega)$ is equal to the complex representation of the output divided by this representation for the input. In this case the absolute value $|G(j\omega)|$ of the complex function $G(j\omega)$ is the ratio of the amplitude of the output sine wave to that of the input. It is then correct to refer to $|G(j\omega)|$ as the magnitude of the quantity $G(j\omega)$ and to use the term magnitude for the amplitude ratio of the sine waves involved. However, no physical system is actually linear. It can only approximate linearity to a greater or less degree. A system may be very nonlinear in the sense that no transfer function exists (in the usual sense) and one may at the same time have a sinusoidal output for a sinusoidal input. We may still form the amplitude ratio for the output and input waves, but this is no longer the magnitude of a transfer function. The use of the term magnitude as it relates to a transfer function is then incorrect. Of course one arbitrarily may associate a complex number with each pair of sine waves so that the amplitude ratio for these waves is the magnitude of this number. However, the amplitude ratio is a "ratio." It therefore appears to the author that it will mean more to the technician if it is referred to as such. We are indebted, no doubt, to electrical engineers for the word magnitude. They are accustomed to thinking in terms of complex numbers, but the average technician is not.

Sine-Wave Generators

A Survey of Pneumatic, Mechanical, and Electrical Devices for Obtaining Frequency-Response Data

By D. W. St. CLAIR,[1] L. W. ERATH,[2] and S. L. GILLESPIE[3]

Data are given on pneumatic, mechanical, and electrical sine-wave generators which are or can be used to impose the "wiggle" necessary for the taking of frequency-response data in the industrial-control field. The high points are covered from a functional rather than a theoretical viewpoint and a reference list is presented to assist those interested in additional information.

Introduction

THE use of frequency-response analysis in the industrial-control field has mushroomed in the past few years. Indeed, data presented here describing test equipment associated with the method, in many instances, will be obsolete before this paper is published.

The purpose of the paper is to assemble information on the various schemes and devices which are or can be used as forcing functions to provide frequency-response test data in the industrial-control field. We believe this information will be valuable to those in the field associated with making frequency-response tests and of interest to a much wider audience.

This survey is of a descriptive rather than analytical nature and will cover only general features of the generators themselves. Other features of the test equipment used for dynamic-data procurement of which the wave generator may be only a part will be mentioned only briefly. We have purposely omitted any lengthy discussion of the benefits to be derived by taking frequency-response data, since the number of generators in use, and the interest in the ASME frequency-response symposium of which this paper is a part, speak for themselves.

While there is some overlapping, an apparatus has been broadly classified as being pneumatic, mechanical, or electrical according to its primary output; this paper has been divided accordingly.

Pneumatic

Inasmuch as pneumatic wave generators are not commercially available (though some are available on special contract), companies wishing to use one have had to develop their own. The most prevalent scheme used to produce a pneumatic wave is first to produce rotary motion, convert to simple harmonic motion, and then to convert to a pneumatic signal. In one instance an electronic oscillator is used to drive an electropneumatic relay. In another, an air compressor is used directly to produce oscillations. At least one pneumatic model has been made using the principle of the relaxation oscillator. Some data also have been received describing a self-excited nozzle-and-baffle arrangement.

Data are assembled in Table 1 on 21 pneumatic sine-wave generators. A few of the column headings need amplification: "Level" represents the limits of the equipment including the amplitude of the sine wave; "dynamic response" refers to 1 psi amplitude dead-ended; "flat" means in the ratio of $1 \div \sqrt{2}$ or vice versa. Wave generators numbered in this table and in the body of the paper may be identified with their owners by reference to the Acknowledgments which have been numbered correspondingly.

The normal arrangement is to use combinations of an electric motor, variable-speed mechanisms, and gears to produce a shaft rotation of the desired frequency. Shaft rotation is converted to simple harmonic motion through an eccentric or Scotch yoke and thence through linkages to a pneumatic device converting position to air pressure. Level and amplitude are set by appropriate linkages.

Those models which vary appreciably from this normal are described in further detail in the following. Additional details which might be of interest also are given according to their number in Table 1.

1 See reference (1).[4]

4 This unit is explosion-proof. It is adapted for generating electrical signals and an electronic oscillator is used when applicable.

5 This model is in addition to one like No. 13.

7 For field work an air motor can be attached. In addition, there is a unit under development which has been field-tested, based on the principle of the relaxation oscillator.

8 This is a new piece of apparatus and more are being planned. Recording is photographic but requires no development, and chart speed is synchronized with frequency which is adjustable in 6 steps by gears. Shaft rotation is converted to an electrical signal through a potentiometer which operates an electropneumatic relay and recorder without amplification. The output from the electropneumatic relay feeds a 3-way valve which introduces the wave to the process but keeps the loop closed.

9 See reference (2). The case has an air or nitrogen purge and the recorder is driven at a speed proportional to frequency. The company has several generators of this description in use. A new generator for electrical outputs is under development.

10 This unit is used for instrument and controller evaluation.

11 This unit is used to test model plants and analogs. Frequency is adjustable in 15 steps; amplitude can be adjusted pneumatically.

12 This model is used for field demonstration. Frequency is adjustable in 6 steps; level is pneumatically adjustable; and a 3-term controller is part of the rig.

13 See reference (3). This wave generator is part of an apparatus for plotting the transfer function directly on polar coordinates. Frequency is controlled by a closed loop and the level

[1] Control Engineer, Kodak Park Works, Eastman Kodak Company, Rochester, N. Y. Assoc. Mem. ASME.

[2] Vice-President, Southwestern Industrial Electronics Company, Houston, Texas.

[3] Design Engineer, Woodward Governor Company, Rockford, Ill.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 26, 1953. Paper No. 53—A-12.

[4] Numbers in parentheses refer to the Bibliography at the end of the paper.

Generator No.	Use		Adjustments											Dynamic Response			Approx. Size		Reproduction Data			Frequency varied by	Type of Recorder	Relay	Type of organization (owner)	Generator and Recorder Combined	Generator No.
			Frequency					Amplitude			Level																
			Range																Man hrs.	Cost of parts in dollars	Cost in dollars						
	Field	Laboratory	Min.	Max.	Continuous	Calibrated	Running	Range Min-Max	Continuous	Calibrated	Range Min-Max	Continuous	Calibrated	Running	Flat in range	Flat to, cps	Peak at, cps	Inches	Lbs.								
1	✓	✓	3CPM	12CPS	✓	X	✓	0-15 PSI	✓	X	1-30 PSI	✓	X	X	X	1	3	10x10x15	30	4	112	125	REVCO	ELECTRONIC	BAILEY	EDUCATION	X	1
2	X	✓	0.1CPH	0.5CPS	✓	X	✓	0-10 PSI	✓	X	0-20PSI	✓	X	✓	✓	1		8x14x20	20	8	215	235	METRON	PNEUMATIC	FOXBORO	CHEMICAL	X	2
3	✓	✓	0.5CPH	18CPS	✓	✓	✓	0-15 PSI	✓	X	0-30 PSI	✓	X	✓	X	5	22	12x18x24	120				GRAHAM	ELECTRONIC	FOXBORO	CHEMICAL	X	3
4	✓	✓	3CPH	15CPS	✓	X	X	0-15 PSI	✓	X	0-30 PSI	✓	X	X		5	5	13x22x28	150	240	300	1500	GRAHAM	ELECTRONIC	MOORE	CHEMICAL	X	4
5	X	✓	3CPH	10CPS	✓	✓	✓	0.3-4 PSI	X	✓	3-20 PSI	✓	✓	X		4	7	10x12x30	60			2000	GRAHAM	ELECTRONIC	NONE	CHEMICAL	X	5
6	✓	✓	0.84CPH	15CPS	✓	X	✓	0.05-5 PSI	✓	✓	0-20 PSI	✓	X	✓	X	9	10	6x10x20	25	40	370	500	METRON	ELECTRONIC	FOXBORO	CHEMICAL	X	6
7	✓	✓	5.7CPH	16CPS	✓	✓	✓	0.2-0.8 PSI	✓	X	3-15 PSI	✓	X	X	X	6	12	12x14x17	50	100			BALL & DISC	ELECTRONIC	FOXBORO	INSTRUMENT	X	7
8	✓	✓	3CPH	0.5CPM	X	✓	✓	0-7 PSI	✓	✓	2-14 PSI	✓	✓	✓				10x18x22	50			1200	GEARS	PHOTOGRAPHIC	HARTMANN & BRAUN	INSTRUMENT	✓	8
9	✓	✓	2CPH	0.5CPS	✓	X	✓	0-6 PSI	✓	✓	3-15 PSI	✓	✓	✓				8x14x22	70		250	1000	D.C. MOTOR	PNEUMATIC		CHEMICAL	✓	9
10	X	✓	10CPM	9CPS.	✓	X	✓	0-6 PSI	✓	✓	3-15 PSI	✓	✓	✓				11x12x13	15	50	175	225	FRICTION DISC.	PNEUMATIC	KENT	INSTRUMENT	X	10
11	X	✓	2.2CPH	3CPM	X	✓	✓	0-6 PSI	✓	✓	3-15 PSI	✓	✓	✓				6x8x14	30	75	250	325	GEARS	PNEUMATIC	KENT	INSTRUMENT	X	11
12	✓	X	3.75CPH	2CPM	X	✓	X	0-6 PSI	✓	✓	3-15 PSI	✓	✓	✓				6x12x15	23	100	300	400	GEARS	PNEUMATIC	KENT	INSTRUMENT	X	12
13	✓	✓	2CPH	12CPS	✓	✓	✓	0.05-7.5 PSI	✓	X	1-19 PSI	✓	✓	✓	✓			11x12x22	85				METRON	ELECTRONIC	MINN-HONEYWELL	INSTRUMENT	X	13
14	✓	✓	0.5CPH	20CPS	✓	✓	✓	0-10 PSI	✓	✓	0-20 PSI	✓	✓	✓		4	7	24x24x24	40			3000	GRAHAM	ELECTRONIC	TAYLOR & SWARTWOUT	PETROLEUM	X	14
15	X	✓	1CPH	15CPS	✓	X	✓		✓	X	✓	✓	X	✓				6x6x13	10				D.C. MOTOR	ELECTRONIC		PETROLEUM		15
16	✓	✓	1CPH	10CPS	X	✓	✓		✓	✓	✓	✓	✓	✓		1							GEARS	PHOTOGRAPHIC		PETROLEUM		16
17	X	✓	3CPH	225 CPS	✓	✓	✓	.0025-2.5PSI	X	✓	3-15 PSI	✓	✓	✓		180	X	12x18x24	50	60	400	650	AUDIO OSC. D.C. MOTOR	ELECTRONIC	TAYLOR	INSTRUMENT	X	17
18	✓	✓	0.13CPH	0.74CPS	X	✓	✓	0.1-6 PSI	✓	✓	0-20 PSI	✓	✓	✓	✓			16x18x24	45	20	255	320	GEARS	PNEUMATIC	TAYLOR	INSTRUMENT	✓	18
19	X	✓	0.2CPS	35CPS	✓	X	✓		X	X	X	0-3000 PSI	✓	✓	✓			14x16x16	25				D.C. MOTOR	ELECTRONIC	NONE	INSTRUMENT	X	19
20	✓	✓	10CPM	250CPS		✓	✓	0-50PSI		✓	0-150 PSI		✓	✓	✓			4x6x10	15	20	50	100	D.C. MOTOR	ELECTRONIC	NONE	GOVERNMENT	X	20
21			100CPS	2000CPS				0-150 PSI			0-125 PSI							8x12x12	15	12	50	75	D.C. MOTOR	ELECTRONIC	NONE	GOVERNMENT	X	21
22		✓		40CPS				0-45°						✓				6x6x6	10			1000				INSTRUMENT	X	22
23	✓	✓	0.3CPM	20CPS	✓	✓	✓	0-1 IN.	✓	✓	✓							6x12x20								PETROLEUM	X	23
24	✓	✓	0.5CPM	18CPS	✓			0-1 IN.		✓	X				✓			8x18x19	187			3000	VICKERS		NONE	INSTRUMENT	X	24

Code: ✓ -Yes; x - No; Blank - No information or not applicable

is adjustable pneumatically. The unit also is equipped to yield electrical and mechanical outputs. A phase-shifting device between the measured signal and a reference also is provided. Data are recorded on a two-pen electronic recorder, are displayed on an oscilloscope, or are plotted automatically on polar co-ordinates.

14 The model described has evolved from three previous models. The apparatus also provides mechanical output (0 to 1.5 in. max) and electrical outputs (0 to 100 max) with the same flexibility of adjustment as shown for air and with responses flat in the range. The maximum mechanical-motion load is set by the 1/4 hp of the Graham drive.

16 This model is in development. It will be explosion-proof and frequency will be changed in 45 steps of about 1:1.25.

17 This unit uses the normal means of obtaining shaft rotations from 3 rph to 10 cps and the shaft drives a sine-wave potentiometer. The output from either this or an electronic oscillator (2 to 225 cps) is fed into a step attenuator (1 and 10-db steps) and then into an amplifier. The output of the amplifier operates an electropneumatic relay.

18 Dimensions and weight given in the chart include pneumatic amplifiers and recorders. Frequency is adjustable in steps of 1:1.6 by a commercially available unit and level can be adjusted pneumatically. Chart speed is synchronized with frequency.

19 This was a limited-purpose generator used for a specific development. Using an air compressor directly, amplitude could be varied by volume changes, level by a reducing valve. Leaks around the piston distorted low frequencies.

20 This generator is a rotary pneumatic valve and operates as a pneumatic push-pull oscillator. The present design exhibits distortions under certain operating conditions but these are being reduced through development.

21 This generator is a siren-type pneumatic oscillator and is a modification of a device reported by R. D. Meyer (4). Distortion is small if consideration is given to proper port design and use is made of a resonant chamber. Amplitude is about 80 per cent of the pressure level.

Information has just been received, briefly describing a device used by the David Taylor Model Basin, Department of the Navy, Carderock, Md., but complete data could not be obtained in time for this survey. The device consists of a vibrating bar, self-excited by a stream of compressed air directed toward the center of the bar by a suitably shaped nozzle. This system tends to go into oscillation at the natural frequency of the bar; frequencies between 800 and 2000 cps have been obtained in connection with calibrating accelerometers.

By and large, the additional comments which have been received may be summarized as follows: The improvements most desired run toward more flexible and precise adjustment of frequency, amplitude, and level. Initial models made for laboratory use are being made explosion-proof for field use. Weight has been considered a possible shortcoming by some in early stages of use but as the tendency toward electronic recording has increased, the problem of test equipment with considerable weight and bulk has been solved with wheels, rather than handles. A rig equipped for adjustment of level pneumatically is uniquely suited for imposing test signals while the control loop under test is closed (or electrical adjustment if the controller is electrical).

The subject of distortion and air-handling capacity is beyond the scope of this paper. Those troubled in this sense may use this paper as a guide to discover who may have a better solution.

Apparent discrepancies between data in articles already published on some of these devices and data presented here are due to more up-to-date information.

Mechanical

There are many mechanisms for obtaining wave motions. If a constant-angular-velocity rotation (5) is to be translated into theoretically perfect linear harmonic motion, five well-known mechanisms which can be used are hypocycloid, Scotch yoke, isosceles sliding block, swash plate, and positive cam.

The conversion of linear oscillation to shaft oscillation is usually accomplished with a rack and pinion. Frequently, however, band-and-wheel construction will allow less intrusion of error.

The conversion between constant angular-velocity rotary motion and harmonic oscillation of a shaft can be accomplished with four-bar linkages, sliders and cranks, and several types of integrators.

Three types of mechanical oscillators in current use or under construction will be described here. One is used in servomechanism studies where the power-output demands are small; a second is an intermediate size and is now under construction for use in a petroleum industry; and a third is of large size for use in the field of speed governors and prime movers. They have been tabulated in this order in Table 1, as Nos. 22, 23, and 24. Additional information is as follows:

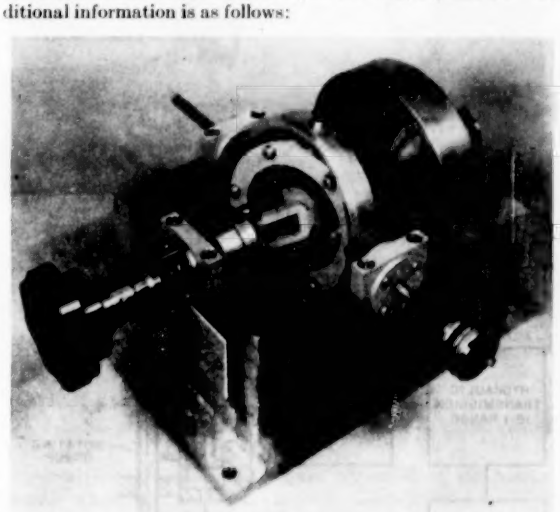
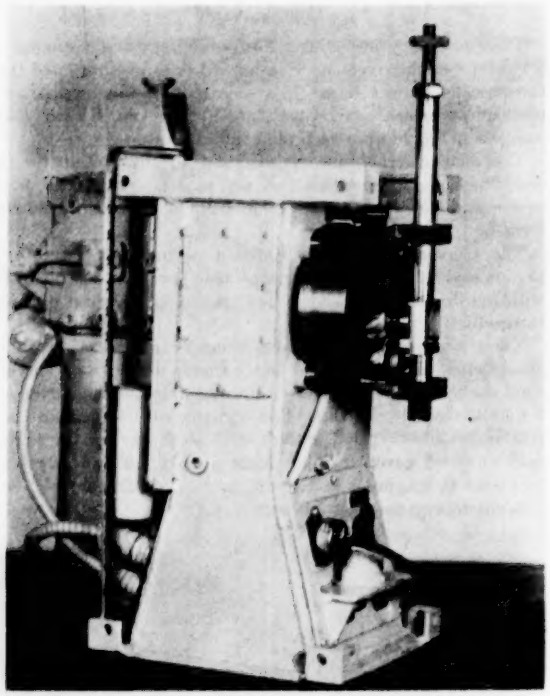


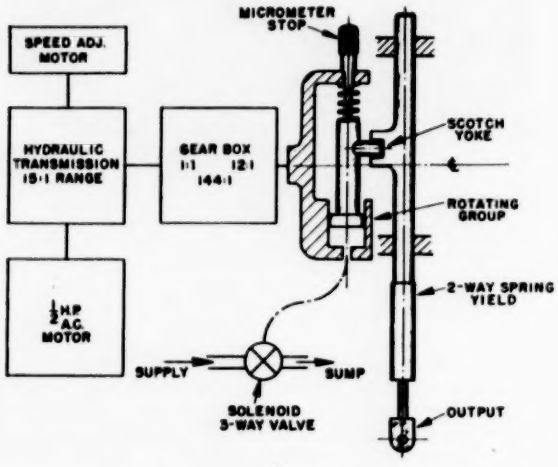
Fig. 1 Device for Obtaining Sinusoidal Angular Displacement From Constant Angular Velocity

22 See reference (6). This unit is shown in Fig. 1. It is built around a large-bore precision ball bearing constrained by a pivot so that the plane of the ball race may be inclined with respect to a plane normal to the shaft axis of the constant-velocity input drive. Rotation of the input shaft produces gyration of the bearing proportional to the angle of inclination in a manner similar to that of a swash plate. The outer race of the bearing is constrained against rotation by a pin connecting it to a ring gimbal. The gimbal ring is supported on a double-ended output shaft which is guided through bearings. Thus as the input shaft revolves the output shaft oscillates through an angle either side of center equal to the angle of bearing-race inclination relative to its shaft. An adjusting knob and mechanism are provided to allow amplitude adjustment while running.

This oscillator can be driven by any convenient constant-speed source (input 2 to 8 oz-in.) and its output applied to mechanisms having a shaft-angle reference input. In one particular arrangement an electronic constant-velocity drive is employed using a two-phase drag-cup induction motor. A gear box between the



(*a*)



(*b*)

FIG. 2 DEVICE FOR OBTAINING HIGH-POWER MECHANICAL OSCILLATIONS

motor and the oscillator gives 0.05 to 1.0 cps and 1.0 to 25 cps. The output goes directly into the servomechanism under study. One user reports a third-harmonic content of 0.25 per cent at 10 deg angular displacement.

23 This unit has just been built and only limited data were obtained in time for this survey. The unit was designed to deliver 3 lb thrust at the maximum stroke and frequency but twice this force has been obtained without difficulty. Also provided is a device to produce an electrical pulse 2 deg wide at each 90-deg interval of sine motion

24 This device, Fig. 2, was designed to be used in determining the frequency response of speed governor-prime mover load combinations.

A schematic diagram of the unit in Fig. 2(*b*) shows the basic arrangement of the device. The crankpin of the Scotch-yoke head is carried on a slide member held against a rigid stop at zero amplitude by a spring. Hydraulic pressure from the Vickers hydraulic transmission is admitted through a solenoid valve to a bellows in the slide. This pressure moves the slide out until it is stopped by a preset micrometer. De-energization of the solenoid valve allows the spring to return the slide quickly to center, a necessary function in the application of the rig.

A spring-yield device is provided to protect the assembly if loads should exceed the design force of 100 lb. The output shaft may be swung through 360 deg for ready alignment with the element to be oscillated; multiple mounting feet also are provided to facilitate anchoring.

In field cases this oscillator has been used in connection with recording equipment, amplifiers, generators, and so on, transported in a mobile test-laboratory truck. A panel in the truck carries all of the controls necessary for the oscillograph operator to start and stop the driving motor, start and stop oscillation, and control frequency over the fifteen-to-one speed range of the transmission. The operator is in telephone communication with an operator located at the oscillator who is provided with a second group of controls arranged so that he can override the controls in the truck. Changes in peak amplitude or shifting of gears must be done with the oscillator shut down.

Electrical

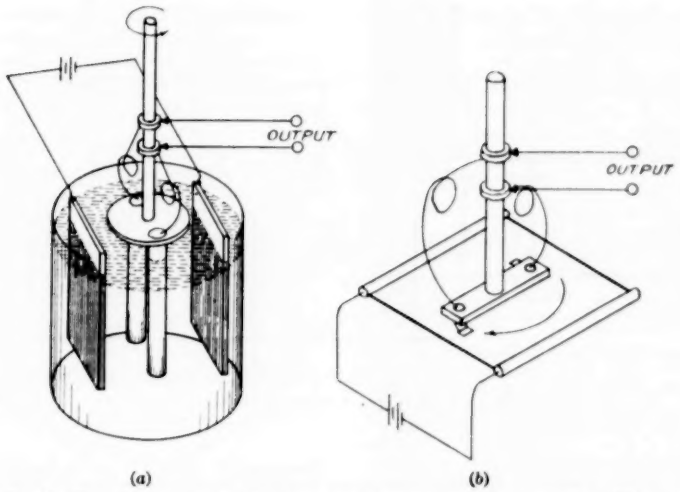
Electric generators may be divided into two general classes; those which use mechanisms and those which do not. The latter are the type commonly referred to as oscillators and are the only variety of wave generator covered in this survey which are commercially available as standard equipment.

Of course any of the methods described previously for obtaining mechanical sine waves can be used to operate a potentiometer or other electrical device. However, a unique electromechanical means of producing low-frequency sine waves is one in which two electrodes or brushes are rotated in a volume of liquid containing flat and parallel equipotential planes or on a surface containing straight and parallel equipotential lines, there being no gradient in either case. Fig. 3(*a* and *b*) illustrates these devices. The limiting factor in this arrangement is, of course, the perfection of the equipotential planes or lines.

Another method of electromechanically generating a sine wave or a wave of almost any shape is the shaped-card potentiometer. An insulating card such as shown in Fig. 4(*a*) is cut to the desired shape, wound with resistance wire, and bent into a circle to be assembled into a potentiometer as shown in Fig. 4(*b*). Rotating the shaft produces the function defined by the card shape.

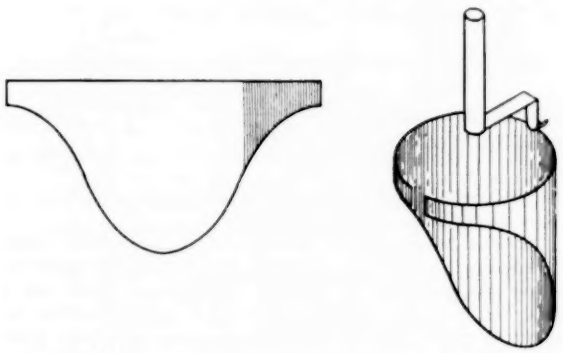
All of the electromechanical generators mentioned have the desirable feature of constant output versus frequency; and when they are kept free of mechanical or electrical reactances there are no transient troubles. They are limited in frequency by the dynamics of their structure, however; and when the generating elements are wire-wound they must have electrical filters to remove the steps which are present in their output voltage.

The other class of sine-wave generators is the oscillator. Hundreds of types of oscillators have been devised and used. Some of the electrical types have been thermally operated and some have made use of the characteristics of rotating electric machines, such as the amplidyne. Vacuum-tube oscillators and signal simulators, are, however, currently enjoying a large popu-



(a) (b)

FIG. 3 TWO EXAMPLES OF A PRINCIPLE FOR OBTAINING ELECTRICAL SINE WAVES



(a) (b)

FIG. 4 THE SHAPED-CARD POTENTIOMETER

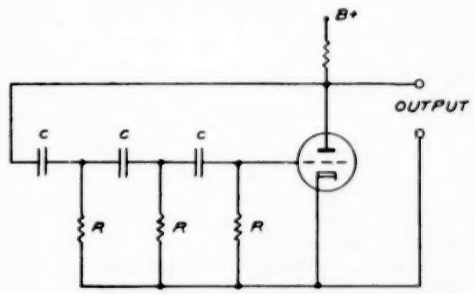


FIG. 5 SIMPLIFIED OSCILLATOR CIRCUIT

larity, and the remainder of the paper will be devoted to this classification.

VACUUM-TUBE OSCILLATORS

One of the earliest forms of vacuum-tube oscillator which was used as a low-frequency generator is known as the phase-shift oscillator and is shown in Fig. 5.

This circuit will oscillate at a frequency for which the phase shift in the resistance-capacitance network is 180 deg, and the amplitude of the oscillation will stabilize when the gain of the amplifying tube is exactly equal to the loss in the network. The frequency can be varied by varying the resistances or capacitances, or both.

Originally, some manual means of controlling the gain was provided and a gain adjustment usually was required for each change of frequency.

In the past few years low-frequency oscillators having automatic level control have appeared in the literature. In general, the time constant of an automatic level-control system must be long compared with the period of oscillation. Oscillators covering frequencies from 1 cps up are quite practical using this system; however, in oscillators intended for use below 1 cps, the long time required for the amplitude to stabilize is annoying and further progress in amplitude control was required before electronic oscillators for very low frequencies could be called practical.

Several circuit arrangements have appeared on the market recently which have made the electronic low-frequency oscillator more practical and have caused its widespread acceptance as a low-frequency signal generator.

It has been possible to arrange the automatic level control so that its time constant is very long when it is acting upon signals of a standard amplitude; however, it becomes short when the generated signal amplitude deviates from this predetermined value. This allows the oscillator to recover rapidly from transients which are generated from switching ranges or from tuning.

Fig. 6 illustrates a modification of the circuit shown in Fig. 5 to incorporate an amplitude-control device of this type. A high-impedance diode detector D_1 generates a d-c voltage across resistance R_1 and capacitance C_1, which is proportional to the level of oscillation. This voltage is then made to control the gain of a variable-gain amplifier which in turn controls the amount of broad-band negative feedback around the amplifier tube T_1. When the level increases, the gain of the variable amplifier is made to increase which reduces the gain of T_1. The level of oscillation is thereby regulated. A second diode D_2 is biased so that it does not normally conduct; however, in the event that the level of oscillation is disturbed by switching ranges and tries to increase, the increased signal is rectified by D_2 which generates a d-c voltage across resistance R_2 and capacitance C_2 (associated with a short time constant) which acts rapidly to reduce the gain

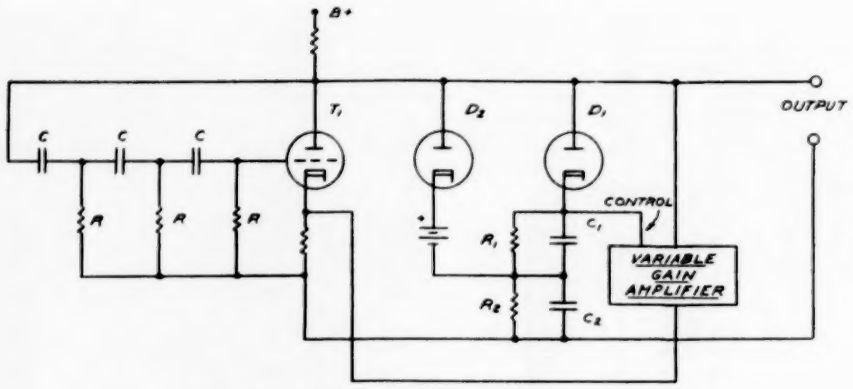


FIG. 6 BASIC DESIGN OF A COMPLEX OSCILLATOR CIRCUIT

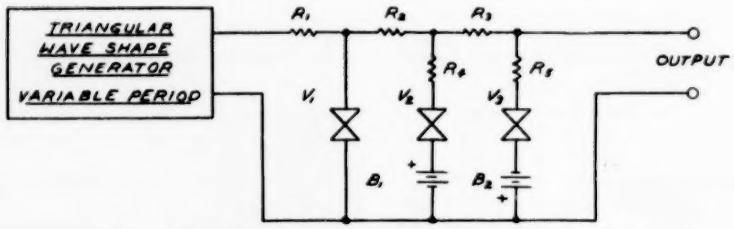


FIG. 7 ANOTHER SCHEME FOR OBTAINING ELECTRICAL OSCILLATIONS

in the T_1. The circuit obviously can be arranged to handle decreases in amplitude as well. For best performance, the excess energy stored in the frequency-determining network also would have to be absorbed. Circuits to accomplish this are not shown. This type of oscillator is capable of producing an excellent wave form.

A second type of low-frequency generator currently available generates a triangular wave which is distorted by a nonlinear network into a sine wave. Fig. 7 illustrates this system. The triangular-wave generator may be any of several well-known circuit arrangements. The most popular arrangement is to generate a square wave with a vacuum-tube flip-flop circuit and then integrate it with a resistance-capacitance network. The nonlinear circuit for distorting the triangular wave in Fig. 7 is a network made up of linear resistances and of varistors. Resistance R_1 and varistor V_1, for example, can be made to round off the peak of the triangular wave, while resistances and varistors R_2, R_4, V_2, R_3, R_5, V_3 in combination with the suitably delayed characteristics produced by the batteries B_1 and B_2, can be adjusted so that the wave is made to approximate closely a sine wave.

The advantage of this system is that the amplitude remains constant when switching or tuning. The obvious disadvantage is that even though the total distortion can be reduced to below 1 per cent rms, the wave shape is lumpy and contains an amount of higher-order harmonics that is sometimes objectionable.

The foregoing discussion has avoided a description of any actual apparatus and has used only simplified circuits for the purpose of aiding in an explanation of the principles involved.

A brief listing of the existing electrical wave generators which have come to our attention during the compilation of this survey follows:

25 The Plessey International Limited has marketed (or is planning to) a "Nyquist Diagram Plotter" which automatically runs through a range of frequencies and plots the results to circular co-ordinates.

26 The Massachusetts Institute of Technology has three generators. In Portable Director Model M-8 low-frequency waves are generated by use of a variable-speed drive and a potentiometer. The potentiometer may be energized either by direct current to produce a low-frequency sine wave or by alternating current to produce a modulated carrier. The Suppressed Carrier Signal Generators D-1 and D-2 generate carrier signals which are modulated at a low frequency by rotating the armature of a synchromotor in an a-c field.

The Hewlett-Packard Company, the Krohn-Hite Instrument Company, and Southwestern Industrial Electronics are three major producers of electronic oscillators. The low-frequency limit of these devices is in the order of 0.01 cps.

The generator shown in Fig. 3(*b*) was available during the last war from Bell Telephone Laboratories but no information was obtained on its current availability.

The potentiometer shown in Fig. 4 is currently available from a large number of suppliers among which are the Technology Instrument Corporation, the Gamewell Company, and the Fairchild Camera and Instrument Corporation.

CONCLUSION

There are many forms of sine-wave generators. While some are considerably more versatile than others it is probably quite safe to say there is no one best piece of equipment because of the great variety of requirements.

The co-authorship of this paper has been carried out in the following manner: D. W. St. Clair—Pneumatic and Co-Ordination; L. W. Erath—Electrical; S. L. Gillespie—Mechanical. Further information on the sine-wave generators may be obtained from the organization listed under the corresponding number in the Acknowledgments. However, such additional informa-

tion is not so available at this time with respect to generators 3, 4, 5, 20, and 21, partly for reasons of national security.

We also take this opportunity to point out that this survey may be made more complete and brought up to date through submitted discussions.

Acknowledgments

We wish to acknowledge the assistance of the following individuals and their organizations for furnishing information and data which have been directly usable in this survey. The numbers correspond to those in Table 1 and in the text:

1 Case Institute of Technology, Cleveland, Ohio, D. P. Eckman.

2 Carbide and Carbon Chemicals Company, South Charleston, W. Va., F. A. Woods.

3 E. I. du Pont de Nemours & Company, Charleston, W. Va., R. S. Lovett.

4 E. I. du Pont de Nemours & Company, Orange, Texas, P. S. Buckley.

5 E. I. du Pont de Nemours & Company, Wilmington, Del., J. T. Thomas.

6 Eastman Kodak Company, Rochester, N. Y., D. W. St. Clair.

7 The Foxboro Company, Foxboro, Mass., A. R. Catheron.

8 Hartmann and Braun, Frankfurt on Main, West Germany, W. Oppelt.

9 Imperial Chemical Industries Ltd., the Frythe, Welwyn, Herts, England, A. R. Aikman.

10, 11, 12 George Kent Ltd., Luton, Bedfordshire, England, R. S. Medock.

13 Minneapolis-Honeywell Regulator Company, Brown Instruments Division, Philadelphia, Pa., E. F. Hochschild.

14 Shell Development Company, Emeryville, Calif., B. D. Stanton. (Unit built by Librascope, Inc., Glendale, Calif. Direct inquiries here.)

15, 16 Shell Laboratorium, Delft, Holland, A. J. Johnson (At Emeryville, Calif.).

17 Taylor Instrument Companies, Rochester, N. Y., E. M. Barr.

18 Taylor Instrument Companies, Rochester, N. Y., L. M. Zoss.

19 Taylor Instrument Companies, Rochester, N. Y., F. B. Newell.

20 United States Naval Air Missile Test Center, Point Mugu, Calif., W. F. Abbott.

21 United States Naval Air Missile Test Center, Point Mugu, Calif., T. E. Hanes.

22 The Ahrendt Instrument Company, College Park, Md., W. R. Ahrendt.

23 Shell Development Company, Emeryville, Calif. (Unit built by Librascope, Inc., Glendale, Calif. Direct inquiries to E. W. Silvertooth here.)

24 Woodward Governor Company, Rockford, Ill., H. B. Capwell.

25 Plessey International Ltd., Ilford, Essex, England, R. W. Hardisty.

26 Massachusetts Institute of Technology, Cambridge, Mass., Emery St. George, Jr.

The following organizations are actively applying frequency-response techniques to evaluations or are in the process of developing wave generators:

Armour Research Foundation of Illinois Institute of Technology, Chicago, Ill., G. A. Nothmann.

Battelle Memorial Institute, Columbus, Ohio, F. C. Todd.

British Iron and Steel Research Association, London, England, S. S. Carlisle.

Elliott Bros. Ltd., Borehamwood, Herts, England, A. E. DeBarr.

We also here express our appreciation to the many individuals who have provided references and given leads in connection with this survey.

BIBLIOGRAPHY

1 "A Pneumatic Sine-Wave Generator for Process Control Study," by D. P. Eckman and J. C. Moise, Instrument Society of America (ISA) Technical Paper 52-2-2, 1952.

2 "A Portable Instrument for Process Control Analysis," by A. R. Aikman, *Instrument Practice* (British publication), May, 1951, p. 393.

3 "An Automatic Transfer Function Measuring and Recording System," by R. J. Ehret, J. M. Embree, E. F. Hochschild, and E. C. Grogan, Proceedings of the AIEE Conference on Recording and Controlling Instruments, Philadelphia, Pa., November, 1952. (Also to be published in AIEE Transactions, *Electrical Engineering*, in digest form.)

4 "Dynamic Pressure Transmitter Calibrator," by R. D. Meyer, *Review of Scientific Instruments*, vol. 17, May, 1946, p. 199.

5 "Signal Generators for Servo System Measurements," by C. F. White, Naval Research Laboratory.

6 "A Frequency Spectrum Analyzer for the Study of Servo Mechanisms," by F. H. Ferguson, Naval Ordnance Laboratory, Memorandum 9317.

7 "Harmonic Response Test Gear," Vickers Armstrong Ltd., Newcastle, England.

8 "A Technique for Rapid Determination of the Harmonic Response Characteristics of a Position Control Servo Mechanism," by E. B. Pearson and G. F. Lingwood, *Instrument Practice*, March, 1953, p. 340.

Discussion

F. A. Woods.[5] The authors are to be congratulated for bringing together so much information on this diversified subject and presenting it in such an interesting manner.

There is little that can be added to the information which has been given on our sine-wave generator itself (generator No. 2 as listed in the paper). However, the methods used by various people in recording and extracting the pertinent data once the sine-wave generator is running on a given test are also of considerable interest.

The measurement of the input and output sine waves produced with the use of this generator is accomplished by means of a two-pen pneumatic strip chart recorder. Two chart speeds are available, 1 iph and 1 ipm, which essentially cover the range of frequencies available. From the chart records obtained the magnitude ratio and phase angle for each frequency tested are extracted and then plotted in the normal manner. This task represents the greatest single drawback to our system since it is quite tedious, and the measurement of phase angle is inclined to be inaccurate for certain frequencies. However, taking all these things into consideration, the system described is simple to operate and relatively inexpensive. It has been found to be quite adequate for laboratory and demonstration purposes.

The writer would be interested in hearing how others have approached and solved these same problems.

E. L. Coulter.[6] Although this paper probably has a good deal of general-interest value, it is especially interesting to those associated with the fields employing frequency-response techniques. To one who is engaged in the use of such equipment in measurements, it offers an unusual opportunity to compare with others who have been faced with the same problems.

[5] Instrument Engineer, Carbide & Carbon Chemicals Company, S. Charleston, W. Va.

[6] Engineer, Air Arm, Westinghouse Electric Corp., Baltimore, Md.

The author is certainly correct when he says that companies wishing to use pneumatic wave generators must develop their own. It is felt that this statement can be extended for mechanical and electrical sine-wave generators also to a large degree. Of course the commercially available equipment does cover the more common electrical and mechanical applications.

Table 1 of the paper is a useful reference both for determining what is being done elsewhere and for comparison.

The two devices which S. L. Gillespie discusses under "mechanical" are novel. They merit general publicity.

L. W. Erath demonstrates a novel application of electrolytics.

A paper of this type is necessarily limited because of the wide usage of the equipment discussed. A broad field of activity is to be found in the aircraft industry. The authors seem generally to have confined themselves to the industrial-control field which also is a very interesting field.

This paper should serve to illustrate the value of papers devoted to surveys. It is felt that other survey papers on other phases of measurement in the frequency-response field could be written.

One method of generating sine waves which was not covered in the paper, and which is believed to be of enough general interest to merit mention, is the so-called underexcited synchro method. This method is treated in the literature[7] but it is felt to be of general enough use to be mentioned in conjunction with such a paper. Briefly, it consists of a continuously rotating underexcited synchro being compared with a normally excited synchro. The resultant signal is a suppressed carrier sine wave modulated at a frequency equal to the speed of rotation of the underexcited synchro. A method of obtaining accurate phase information at high servo frequencies is available by using a third synchro and a mechanical differential.

[7] "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley and Sons, Inc., New York, N. Y., 1948.

It is easily understandable that with three experts such as the authors, the paper would naturally divide into three subdivisions, one for the specialty of each. It is hoped, however, that the large amount of overlap between the three fields is not obscured by this division. In fact, attempts at building universal sine-wave generators have been made and a fair degree of success has been achieved in some cases. One particular model of this type was developed by the writer's company.

In order to demonstrate the degree of overlap of the three divisions a brief description of this generator will be given. The speed range is from 0.005 to 54 cps. This range is obtained with a closed-loop magnetic-amplifier speed control on a servomotor and four gear ratios.

The output of the speed control is a continuously rotating shaft. Driven by this shaft are a Scotch yoke and two continuously rotating synchros which may be adjusted differentially. The Scotch yoke drives a rack and pinion which in turn oscillates a shaft on which potentiometers and synchros are mounted. This shaft is also available for rotational harmonic motion and the rack is available for rectilinear harmonic motion. If the electrical output is used to drive a transfer valve which actuates a bellows the device can become pneumatic also.

It is hoped that this discussion demonstrates that one good set of measuring apparatus may be adapted for use in most applications. The amount of specialization should depend on the amount of variation of jobs handled by the equipment user. Rarely would one user require three different sets of equipment.

Authors' Closure

The authors thank each of the discussers for his interest in the survey. Mr. Coulter is quite correct when he points out the great amount of overlap between the fields. While we have no specific figures it appears that the wave generators being used the greatest number of hours per day on industrial applications are those combining two and three fields of operation.

A Bibliography of the Frequency-Response Method as Applied to Automatic-Feedback-Control Systems

By A. M. FUCHS,[1] TETERBORO, N. J.

The frequency-response method has grown from its beginning during World War II to its present position as the principal method of analyzing and synthesizing automatic-feedback-control systems. This growth can be traced by an examination of the pertinent published material. Dividing this material into fields of interest, such as, Basic Theory; Nonlinear Systems; Books; Applications—Autopilot, Process Control, Components, and so on, makes it possible to study the development of these diverse aspects of the frequency-response method.

Introduction

THIS Bibliography has been prepared with the following objectives in mind:

1 To offer a listing of the major contributions to the theory of the frequency-response method and of its application in the analysis, synthesis, and testing of automatic-feedback-control systems.

2 To present this material in a manner that will make it more accessible to the engineer who consults it to find the solution for a particular problem.

Limitations on the time which could be applied to assembling the publications for this Bibliography have made it impossible to make it absolutely complete.

Since the Bibliography has been restricted to the frequency-response method, references to other broad fields of automatic-feedback-control engineering have been omitted. Reference (284) covers these other fields as well as the frequency-response method, and is recommended for those interested in this broader scope.

As an additional consequence of limiting the referenced publications to the frequency-response method, an element of decision arises as to whether the particular publication falls within the scope of this Bibliography. In those sections of the Bibliography dealing with the theory of the frequency-response method, the decisions have been made rather rigorously. Either frequency-response method was clearly indicated or the article was omitted. In the application sections of the Bibliography, however, the policy was considerably more flexible. Application papers which implied only the use of the frequency-response method have been included. This was done to give the reader a broader view of the tremendous range of activities to which the frequency-response method has been applied.

It is the author's sincere hope that no one will feel slighted by an omission due to oversight or the overzealous application of the foregoing criterion.

The Bibliography is first divided into "fields of interest." They are presented in an order which permits a logical development of the over-all subject matter.

The larger fields of interest are subdivided chronologically by years, and within each year alphabetically by the author's name. Additional listings by title and by author's name in alphabetical order are also available.[2]

The chronological listing within a field of interest does not imply credit for precedence in the discovery or application of the frequency-response method. It is only meant to show the general flow of thinking and achievement within that field of interest, and to facilitate the selection of the desired publications.

Discussion

The discussion of the individual publications which follows is restricted to those items with which the author is familar. To provide a more detailed quantitative report would require more space than has been allotted and the combined efforts of a group of specialists from each particular field of interest.

Prior to the development of the frequency-response method there were many significant contributions to control-system engineering (references 1 to 12 and bibliography of 163). Minorsky's work (1, 2)[3] is interesting for its thorough and original analysis followed by a successful application to the holding of a battleship on course His prophecies of the future of automatic control systems and radio techniques in sea warfare are startlingly accurate.

Hazen's work (3, 4, 7) included the first analytic treatment of instrument servomechanisms. He defined a servomechanism and designed an electric motor for optimum servomechanism performance.

A clear idea of the limitations of the design procedures which preceded the frequency-response method can be found in references (11, 20) and (168, pp. 156-157). The most complete bibliography of this period is found in reference (160).

The theoretical background for the frequency-response method was developed by Nyquist (13), Black (14), and Bode (15). Nyquist established the relationships between open-loop frequency response of a system and its closed-loop stability. Black employed this work to investigate the limitations it placed upon the design of feedback amplifiers and Bode demonstrated the necessary relationships between the amplitude and phase characteristics of realizable systems.

Operational mathematics was developed as an analysis tool (16, 17), and is today one of the essential cornerstones to a control-engineer's education.

Under the impetus of World War II, large groups of engineers at M.I.T. (including Hall, Brown, Campbell, Edwards, Nichols,

[1] Senior Project Engineer, Automatic Controls Laboratory, Eclipse-Pioneer Division, Bendix Aviation Corporation.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 23, 1953. Paper No. 53—A-13.

[2] Write to the author, Automatic Controls Laboratory, Eclipse-Pioneer, Teterboro, N. J.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

Forrester, Chance, and Phillips) and at Bell Telephone Labs (Nyquist, Bode, Black, Ferrell, Dietzold, and MacColl) worked on the problems of employing servomechanisms as a military weapon. Harris' paper (20) was the first to suggest the use of the frequency-response method while Hall's[4] work (21) established the frequency-response method in almost every important aspect of its present application to the analysis and synthesis of linear systems.

Security restrictions limited the circulation of Hall's work until the end of World War II. At that time a large number of significant papers were published (23, 26, 27, 28, 29) which outlined the wartime development of the frequency-response method.

Marcy (30), Harris (38), and Criss (35) outlined the use of the inverse Nyquist plane in analyzing servomechanisms. Other workers (23, 165, 168, 169) revealed the use of logarithmic coordinates as a valuable tool in reducing the labor involved in studying servomechanisms by the frequency-response method.

The methods of compensation of servomechanisms were explored at great length (33, 38, 45, 60). As the advantage of a-c servomechanisms (use a carrier rather than direct current) became evident, their special compensation problems were studied (54, 68, 71, 75, 76, 81).

There was a constant need to reassess the relationship between the frequency response of the system and its transient behavior as shown in references (55, 89, 90, 97). Frequently there arose the need to re-express Nyquist's stability criterion either for a better understanding, an easier method of application, or to present it to a new group of engineers (66, 69, 79, 87).

As the design of linear systems progressed, it became evident that the nonlinearities in almost every control system were one of the most serious sources of malfunctioning. Two principal techniques for handling this problem were developed; one was an analytic approach (126–162) and the other was by using computers (107–125) to simulate the system and its nonlinearities. The computer approach also was stimulated by the application of servomechanisms to ever more complex systems in which more conventional analytical techniques were too lengthy (see Application-Autopilot).

Analog computers were employed to study automatic feedback-control systems prior to the extensive use of the frequency-response method (107, 108, 109). A full understanding of the significance of the frequency-response method gave a new and valuable tool which made the computer approach to the study of control-system performance more fruitful.

The principal advantage of the analog-computer study of control systems is the relative ease and speed with which many control-system configurations may be examined. Furthermore, in dealing with nonlinearities, it by-passes the need for an over-all mathematical solution. The computer requires only that the transfer function of the nonlinear element be reproducible. With modern electronic techniques, this is almost always possible.

Examples of the application of analog computers to the study of nonlinear systems is given in reference (112) (coulomb friction), reference (121) (nonlinear damping), reference (123) (saturation), reference (123) (autopilot nonlinearities). The construction of several computers is found in references (107, 114, 115, 116, 125). Hall's article (115) is particularly interesting in showing the large part which precision-instrument servomechanisms play in the modern, large-scale, analog computer.

Many useful examples exist (126–134) of analytic approaches to nonlinear systems: Kochenburger's work (138) permitted a straightforward modification in the frequency-response method to handle a contactor-type servomechanism. The new analytic technique is called the "describing-function" method. Johnson's paper (152) extended the describing function to numerous other nonlinearities. He included a method for evaluating the error involved in making the approximations required by the describing-function method.

Linvill (144, 161), Ragazzini and Zadeh (157) found a method of handling sampled-data control systems which also requires only a slight modification of the frequency-response method.

On the surface it appears that much has been done to make possible the straightforward analysis of nonlinear servomechanisms. It is primarily in the realm of synthesis of nonlinear systems, in which nonlinearities are intentionally added to improve the system performance, that much remains to be done (154, 138).

As the use of the frequency-response method spread, many books were published to give the student and the engineer a complete background in the field.

References (168) and (172) (Radiation Laboratory Series vols. 25 and 21) present the theory and the application of the frequency-response method as evolved at the Radiation Laboratory, M.I.T., during World War II.

Reference (171) (Brown and Campbell) is the first complete textbook in the field. It has found widespread use in the colleges of our country. It has recently been supplemented by reference (178) (Ahrendt and Taplin), reference (179) (Chestnut and Mayer, vol. 1), and reference (184) (Thaler and Brown). Each of these books provides the student with a background in "operational mathematics" before continuing on through the frequency-response method and its application.

Special mention must be made of reference (183) which covers the papers presented at the Cranfield Conference. A number of these papers are individually listed (83, 87, 92, 145, 155, 156, 189, 190, 224, 233, 265). They offer a broad picture of most of the recent developments in automatic feedback-control engineering.

The application of the frequency-response method to the process-control field has been extensive. A number of the publications which relate the frequency-response method to the process-control field are listed (185–196). Reference (196) gives a more comprehensive bibliography of this field as well as a discussion of the significance of the developments in our ability to handle process-control problems.

High performance of aircraft has often resulted in unstable flight characteristics. Automatic-feedback-control systems are employed to correct this condition. They permit the pilot to maneuver the plane with a minimum of work. As the number of functions which these aircraft-control systems handled increased, it finally became possible to have the plane fly completely automatically. The over-all automatic flight-control system became known as an autopilot.

References (197–208) present the development of aircraft-control systems and autopilots by the use of the frequency-response system. Bollay's article (204) gives a thorough treatment of the entire field of aircraft controls. Much of the theoretical development of frequency-response theory was done with an aircraft-control problem as its basic motivation.

The use of the frequency-response method requires special test equipment to obtain the frequency-response characteristic of the component or system under test. References (209–218) and (171, chapter 10) suggest numerous ways by which this characteristic can be obtained. In making frequency-response measurements, a "director" is employed to generate the input test function. A "breadboard" can be used to permit electromechanical assemblies to be put together and tested rapidly. Both of these items are commercially available from a number of sources.

With the development and application of the frequency-response method has come the design of a wide variety of components especially suited to servomechanism applications.

[4] Reference (18) is an extremely early reference to the frequency-response method, but it had a high security classification and therefore a low circulation.

At high power levels where small size and high speed of response are important, hydraulic mechanisms have played a large part (219 to 229). References (220, 222, 224) describe the applications; references (225, 227, 229) discuss the design problems; and references (221, 226, 228) outline the performance characteristics which are associated with hydraulic components.

At the other end of the power scale are selsyns (237–242) which are used in remote positioning and indicating instruments.

Small motors of both the d-c (244) and a-c type (245, 246) have been designed for servomechanism application.

In the medium power range the amplidyne has been used. Its characteristics are given in reference (249) and applications are given in references (250–252).

An unusual example of the extent to which component analysis is carried when precision feedback units (in this instance a tachometer) are analyzed is given by Frazier (248). It includes the derivation of the delays present in the tachometer.

Several examples of the application of the frequency-response method to industrial problems and military uses are listed. The applications discussed in these references represent only a small part of the use of the frequency-response method in these fields. Many industrial applications are never published, or if published, do not include the frequency-response analysis of the system. The tremendous range of military application of the frequency-response method is largely covered in restricted publications or internal reports of the many research and development organizations in our country. Hence they are not available to this Bibliography.

Using control systems which include a human operator as an element in the control loop, has brought the frequency-response method into the analysis of human-behavior characteristics. References (230–236) give examples of the attempts made to characterize the human operator as an element in a control system. Man being a highly nonlinear system, these efforts have met with only limited success.

As with all fields, there has been difficulty in keeping all the different engineering groups who employ automatic feedback-control systems speaking the same language. Even within a particular engineering society differences arise as to correct terminology. References (279–283) give some of the efforts made to unify the terminology of automatic feedback-control systems.

Finally, there is the problem of educating engineers in the new field of control engineering. Brown and Campbell (262) have considered this problem. One can expect to hear much more on this subject as automatic-feedback-control engineering becomes an essential part of every engineer's background.

Acknowledgment

The author wishes to thank Miss M. Baker of the Eclipse-Pioneer Library, for her assistance in the preparation of the Bibliography; Mrs. M. Turner for the typing; and Mr. J D. Peterson for his encouragement of this work.

BIBLIOGRAPHY

Significant Contributions to Control Before Introduction of Frequency-Response Method

1 "Directional Stability of Automatically Steered Bodies," by N. Minorsky, *Journal of the American Society of Naval Engineers*, vol. 34, May, 1922.

2 "Automatic Steering Tests," by N. Minorsky, *Journal of the American Society of Naval Engineers*, vol. 42, May, 1930.

3 "Theory of Servomechanisms," by H. L. Hazen, *Journal of The Franklin Institute*, vol. 218, September, 1934, pp. 279–330.

4 "Design and Test of a High-Performance Servomechanism," by H. L. Hazen, *Journal of The Franklin Institute*, vol. 218, November, 1934, pp. 543–580.

5 "Principles Underlying the Rational Solution of Automatic Control Problems," by S. D. Mitireff, Trans. ASME, vol. 57, 1935. pp. 159–163.

6 "Time Lag in a Control System," by D. R. Hartree, A. Porter, and A. Callender, Trans. of the Royal Society of London England, series A, vol. 235, 1936, pp. 415–444.

7 "An Automatic Curve Follower," by H. L. Hazen, J. J. Jaeger, and G. S. Brown, *Review of Scientific Instruments*, vol. 7, September, 1936.

8 "Time Lag in a Control System, II," by D. R. Hartree, A. Porter, A. Callender, and A. B. Stevenson, Proceedings of the Royal Society of London, England, series A, vol. 161, 1937, pp. 460–476.

9 "Application of Temperature Controllers," by E. D. Haigler, Trans. ASME, vol. 60, 1938, pp. 633–640.

10 "Constant Speed Control Theory," by H. K. Weiss, *Journal of the Aeronautical Sciences*, vol. 61, February, 1939.

11 "Behavior and Design of Servomechanisms," by G. S. Brown, printed November, 1940, under the Auspices of the Fire Control Commission (See D-2) of NDRC.

12 "Automatic Control in the Presence of Process Lags," by C. E. Mason and G. A. Philbrick, Trans. ASME, vol. 62, 1940, pp. 295–309.

Theoretical Background Material

13 "Regeneration Theory," by H. Nyquist, *Bell System Technical Journal*, vol. 11, January, 1932, pp. 126–147.

14 "Stabilized Feedback Amplifiers," by H. S. Black, *Bell System Technical Journal*, vol. 13, January, 1934, pp. 1–18.

15 "Relations Between Attenuation and Phase in Feedback Amplifier Design," by H. W. Bode, *Bell System Technical Journal*, vol. 19, July, 1940, pp. 421–454.

16 "Transients in Linear Systems," by M. F. Gardner and J. L. Barnes, John Wiley & Sons, Inc., New York, N. Y., 1942.

17 "Modern Operational Mathematics in Engineering," by R. V. Churchill, McGraw-Hill Book Company, Inc., New York, N. Y., 1944.

Basic Theory

1941

18 "Stabilization of Servomechanisms," by D. C. Bomberger and B. T. Weber, Bell Telephone Laboratory, M.M-41-110-52, December, 10.

1942

19 "Fundamental Studies in Servomechanisms Rated Approximately 100 Watts," by J. W. Forester, E. St. George, D. P. Campbell, R. R. Everett, and A. C. Hall, Servomechanism Laboratory, M.I.T., Division of Industrial Cooperation, Project No. 6097, Government Contract No. W 241.

20 "The Analysis and Design of Servomechanisms," by H. J. Harris, Printed under auspices of the Fire Control Committee (Sec. D-2) of NDRC, OSRD, No. 454.

1943

21 "Analysis and Synthesis of Linear Servomechanisms" (Bib.), by A. C. Hall, Technology Press, Cambridge, Mass.

1944

22 "Contributions to the Theory of Automatic Controllers and Followers," (Bib.), by D. G. Prinz, *Journal of Scientific Instruments*, vol. 21, pp. 53–64, 110–111.

1945

23 "The Servo Problem as a Transmission Problem," by E. B. Ferrell, Proceedings of the IRE, vol. 33, pp. 763–767.

24 "A Survey of Criteria for Stability," by D. G. Prinz, Proceedings of the Institution of Chemical Engineers, vol. 27.

25 "Stabilizing a Suction-Relief Valve" (see Discussion), by E. S. Smith, Trans. ASME, vol. 67, p. 87.

1946

26 "Dynamic Behavior and Design of Servomechanisms" (Bib.), by G. S. Brown and A. C. Hall, Trans. ASME, vol. 68, pp. 503–524.

27 "Linear Servo Theory," by R. E. Graham, *Bell System Technical Journal*, vol. 25, pp. 616–651.

28 "Application of Circuit Theory to the Design of Servomechanisms," by A. C. Hall, *Journal of The Franklin Institute*, vol. 242, pp. 279–307.

29 "Frequency Response of Automatic Control Systems," by H. Harris, Jr., Trans. AIEE, vol. 65, pp. 539–545.

30 "Parallel Circuits in Servomechanisms," by H. Tyler Marcy, Trans. AIEE, vol. 65, pp. 521–529.

31 "Analysis of Servomechanisms for Instrument and Power Drives," by G. S. Perkins, *Product Engineering*, vol. 17, pp. 332–334.

32 "Theory of Servo Systems With Particular Reference to

Stabilization," by A. L. Whiteley, *Journal of the IEE*, vol. 93, part 2, pp. 353–367.

1947

33 "Obtaining Attenuation Frequency Characteristics for Servomechanisms," by H. Chestnut, *General Electric Review*, December, pp. 38–44.

34 "A New Method for Assessing the Stability of Linear Regulating Systems" (in German), by L. Z. Cremer, *Math. Mech.*, vol. 26/27, August–September, pp. 161–163.

35 "Inverse Nyquist Plane in Servomechanism Theory," by G. B. Criss, Proceedings of the IRE, vol. 35, pp. 1503–1504.

36 "Theoretical and Experimental Methods for the Study of Closed Clyde Automatic Control Systems," by L. van Hamos, Trans. Instrument Meas. Conf., Stockholm, Sweden, pp. 129–138.

37 "The Analysis and Design of Servomechanisms," by H. Harris, Jr., Trans. ASME, vol. 69, pp. 267–280.

38 "A Comparison of Two Basic Servomechanisms Types," by H. Harris, Jr., Trans. AIEE, vol. 66, pp. 83–93.

39 "Elements of Position Control," by K. A. Hayes, *Journal of the IEE*, vol. 94, part 2A, pp. 161–176.

40 "The Analysis and an Optimum Synthesis of Linear Servomechanisms," by Donald Herr and Irving Gest, Trans. AIEE, vol. 66, pp. 959–970.

41 "Dimensionless Analysis of Angular-Position Servomechanisms," by S. W. Herwald and G. D. McCann, Trans. AIEE, vol. 65, pp. 636–639.

42 "Recent Progress in the Determination of Precision and Stability of Servomechanisms" (in French), by G. Lehmann, Bulletin of the Technical Marit. Aero., vol. 46, pp. 79–110; also Memor. Artill. fr., vol. 21, pp. 505–536.

43 "Progress in Dynamic Stability and Control Research," by W. F. Milliken, Jr., *Journal of the Aeronautical Sciences*, vol. 14 (Bib.), pp. 493–519.

44 "The Frequency-Response Method of Control Analysis," by W. C. Moog, *Instruments*, vol. 20, pp. 810–816.

45 "The Application of Lead Networks and Sinusoidal Analysis to Automatic Control Systems," by G. Schwartz, Trans. AIEE, vol. 66, pp. 69–77.

46 "The Frequency-Response Method in the Theory of Regulation," by V. V. Solodovnikov, *Avtomatika i Telemekhanika*, vol. 8, pp. 65–88.

47 "Properties of Electrical Servo-Systems and Automatic Regulators by Means of Frequency Characteristics" (in Russian), by V. V. Solodovnikov, *Elektrichestuo*, pp. 22–28.

48 "Fundamental Principles of Automatic Regulators and Servomechanisms," by A. L. Whiteley, *Journal of the IEE*, vol. 94, part 2A, pp. 5–22.

1948

49 "A Classification of Linear Transfer Members," by H. L. Mason, Trans. ASME, vol. 71, 1949, pp. 407–412.

50 "Stabilizing Servomechanisms," by Donald McDonald, *Electronics*, vol. 21, pp. 112–116.

51 "Recent Methods for Investigating the Stability of Regulating-Processes" (in German), by W. Oppelt, *Technik*, vol. 3, pp. 312–314.

52 "Basic Principles of Automatic Control Systems," by A. Porter, Proceedings of The Institution of Mechanical Engineers, vol. 159, pp. 25–45.

53 "Application of the Performance Operator to Aircraft Automatic Control," by R. C. Seamons, B. G. Bromberg, and L. E. Payne, *Journal of the Aeronautical Sciences*, vol. 15, pp. 535–555.

54 "Stabilization of Carrier Frequency Servomechanisms," by A. Sobaczyk, *Journal of The Franklin Institute*, vol. 246, July-August-September, pp. 21–43, 95–121, 187–213.

Basic Theory

1949

55 "Comparison of Steady-State and Transient Performance of Servomechanisms," by H. Chestnut and R. W. Mayer, Trans. AIEE, vol. 68, pp. 765–777.

56 "Automatic Control" (in Dutch), by C. Dellaert, *Ingenieur*, vol. 61, September, pp. 63–69.

57 "A New Harmonic Method of Studying the Stability of Linear Systems" (in French), by M. Demontoignier and P. Lefevre, *General Electric Review*, vol. 58, pp. 263–279.

58 "Graphical Determination of Transfer Function Loci for Servomechanism Components and Systems," by E. C. Easton and C. H. Thomas, Trans. AIEE, vol. 68, part I, pp. 307–318.

59 "Servomechanisms and Regulators; Criteria of Stability; Examples of Applications" (in French), by C. Galmiche, *General Electric Review*, vol. 58, January, pp. 19–30.

60 "Damper-Stabilized Instrument Servomechanisms," by A. C. Hall, Trans. AIEE, vol. 68, part I, pp. 299–306.

61 "Equations for the Approximate Solution of Dynamic Problems for Stable Linear Systems," by K. H. Hohenemser, *Journal of the Aeronautical Sciences*, vol. 16, pp. 723–728.

62 "Methods for Determining Frequency Response of Engines and Control Systems From Transient Data," by M. E. LaVerne and A. S. Boseabom, NACA TN 1395, August.

63 "Calculation of the Optimum Parameters for a Following System," by C. Mack, *Philosophical Magazine*, London, England, vol. 40, pp. 922–928.

64 "Electromechanical Lead Networks for A. C. Servo Mechanisms," by Donald McDonald, *Review of the Scientific Institute*, vol. 20, pp. 775–779.

65 "The Inverse Derivative—A New Method of Automatic Control," by C. B. Moore, *Instruments*, vol. 22, pp. 216–219.

66 "A Generalization of Nyquist's Stability Criteria," by A. Vazsonyi, *Journal of Applied Physics*, vol. 20, pp. 863–867.

67 "The Stability of Automatic Regulation" (in Swedish), by E. Voipo, Kraft. o. Ljus, vol. 22, pp. 247–265.

1950

68 "Carrier Compensation for Servomechanisms" (Bib.), by H. E. Blanton, *Journal of The Franklin Institute*, vol. 250, November–December, pp. 391–407, 525–542.

69 "Nyquist Diagrams and Routh-Hurwitz Stability Criterion," by F. E. Bothwell, Proceedings of the IRE, vol. 38, pp. 1345–1348.

70 "Modern Electric Automatic Control Systems" (in Italian), by C. Constadoni, *Elettrotecnica*, vol. 57, March 10–25, pp. 82–95; April 12, pp. 126–148.

71 "Integral Controller for Use in Carrier-Type Servomechanisms," by R. L. Cosgriff, Trans. AIEE, vol. 69, pp. 1379–1383.

72 "Servo Control Problems," by A. J. O. Cruickshank, Trans. Institution of Engineers and Shipbuilders, Scotland, vol. 94, part 5, pp. 266–302.

73 "Methods for the Study of Servomechanisms" (in Italian), by F. Guarnaschelli, *Aerotecnica*, vol. 30, December, pp. 316–322.

74 "Note on the Maximum Precision of Perfectly Stable Servomechanisms" (in French), by G. Lehmann, *Onde Elect.*, vol. 30, June, pp. 267–270.

75 "Improvements in the Characteristics of A. C. Lead Networks for Servomechanisms," by Donald McDonald, Trans. AIEE, vol. 69, part I, pp. 293–300.

76 "Phase Lead for A. C. Servo Systems With Compensation for Carrier Frequency Changes," by A. P. Notthoff, Jr., Trans. AIEE, vol. 69, part I, pp. 285–291.

77 "Determination of Frequency Characteristics From Response to Arbitrary Input," by L. B. Rea and E. R. Walters, *Journal of the Aeronautical Sciences*, vol. 17, pp. 446–452.

78 "Feedback Amplifiers and Servo Systems," by E. E. Ward, *Wireless World*, vol. 27, May, pp. 146–153.

79 "The Nyquist Criterion of Stability," by J. C. West, *Electronic Eng.*, vol. 22, May, pp. 169–172.

Basic Theory

1951

80 "The Frequency Response Approach to Automatic Control Problems," by A. R. Aikman, Trans. Society of Instrument Technology, vol. 3, pp. 2–16.

81 "Effects of Carrier Shifts on Derivative Networks for A. C. Servomechanisms," by G. M. Attura, Trans. AIEE, vol. 70, part II, pp. 612–618.

82 "Network Synthesis by Graphical Methods for A. C. Servomechanisms," by G. A. Bjornson, Trans. AIEE, vol. 70, part I, pp. 619–625.

83 "The Application of the Electrolytic Tank to Servo-Mechanisms Design," by A. R. Boothroyd and J. H. Westcott, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 87–102.

84 "Stabilization of Continuous-Current Servomechanisms" (in French), by M. Cambornac and F. Lajeunesse, *Onde. Elec.*, vol. 31, pp. 434–445.

85 "On the Automatic Control of Generalized Passive Systems," by Kun Li Chien, J. A. Hrones, and J. B. Reswick, Trans. ASME, vol. 74, 1952, pp. 175–185.

86 "Regulation Theory" (in Danish)—I—"Direct Regulation," II—"Regulation With a Servomotor," by P. Draminsky, *Ingenioren*, vol. 60, September 1, pp. 602–606; September 8, pp. 614–619.

87 "The Stability Criterion," by L. S. Dzang, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 13–24.

88 "Survey of the Development in Regulation Technique in Recent Years" (in Swedish), by A. Garde, *Ingenioren*, vol. 60, pp. 549–557.

89 "Servomechanism Transient Performance From Decibel-Log Frequency Plots," by H. Harris, Jr., M. J. Kirby, and E. F. Von Aix, Trans. AIEE, vol. 70, part II, pp. 1452–1459.

90 "Importance of Extending Nyquist Servomechanism Analysis to Include Transient Response," by R. L. Johnson and J. B. Rea, *Journal of the Aeronautical Sciences*, vol. 18, pp. 43–49.

91 "Selection of Optimum Parameters of Automatic Regulators and Servomechanisms" (in Russian), by V. S. Kulebakin, Dokl. Akad. Nauk, SSSR, vol. 77, pp. 265–267.

92 "A Generalization of the Frequency Response Method for the Study of Feedback Control Systems," by N. L. Kusters and W. J. M. Moore, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 105–119.

93 "Combination Open-Cycle, Closed-Cycle Control Systems," by J. R. Moore, Proceedings of the IRE, vol. 39, pp. 1421–1432.

94 "A Theoretical and Experimental Method of Modulation Analysis for the Design of A. C. Servo Systems," by D. Morris, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 521–532.

95 "A Note on the Design of Conditionally Stable Feedback Systems," by Paul Travers, Trans. AIEE, vol. 70, part I, pp. 626–630.

96 "A New Parallel-Type Compensation Network," by G. F. Warnke and V. H. Disney, AIEE Miscellaneous Paper 52–95, December.

97 "A Simple Connection Between Closed-Loop Transient Response and Open-Loop Frequency Response," by J. C. West and J. Potts, Proceedings of the IEE, London, England, vol. 100, part 2, pp. 201–208.

98 "The Development of Relationships Concerning the Frequency Bandwidth and the Mean Square Error of Servo Systems From Properties of Gain-Frequency Characteristics," by J. H. Westcott, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 45–60.

99 "Servomechanisms: A Review of Progress" (Bib.), by A. L. Whiteley, Proceedings of the IEE, London, England, vol. 98, part 1, pp. 289–297; also *Engineering*, vol. 172, December 28, pp. 825–826.

Basic Theory

1952

100 "Logarithmic Analysis of Servo Stability," by N. Alpert, *Product Engineering*, vol. 23, April, pp. 186–190.

101 "Signal Component Control," by D. J. Gimpel and J. F. Calvert, Trans. AIEE, vol. 71, part II, pp. 337–343.

102 "A Theory of Multidimensional Servo Systems," by M. Golomb and E. Usdis, *Journal of The Franklin Institute*, vol. 253, January, pp. 29–57.

103 "A Generalized Method for Analyzing Servomechanisms," by A. A. Hauser, Jr., Proceedings of the IRE, vol. 40, February, pp. 197–202.

104 "Considerations on Secondary Loops in Regulated Systems" (in French), by P. Naslin, Mem. Artill, fr., vol. 26, pp. 59–119.

105 "Stabilization Templates for Servomechanisms," by J. M. Smith, Trans. AIEE, vol. 71, part II, pp. 220–224.

106 "Step Response as a Short Cut to Frequency Response," by D. W. St. Clair, ISA Technical Paper 52-9-3, September.

Analog Computers and Their Applications to Linear and Nonlinear Servomechanisms Studies

1945

107 "A New Type of Differential Analyzer," by V. Bush and S. H. Caldwell, *Journal of The Franklin Institute*, vol. 240, October, pp. 255–326.

108 "Electrical-Analogy Method for Fundamental Investigations in Automatic Control," by D. P. Eckman and W. H. Wannamaker, Trans. ASME, vol. 67, pp. 81–86.

1946

109 "Electrical Analogy Methods Applied to Servomechanism Problems"—Part I; "Dimensionless Analysis of Servomechanisms by Electrical Analogy"—Part II, by G. D. McCann, S. W. Herwald, and H. S. Kirschbaum, Trans. AIEE, vol. 65, part I, February, pp. 91–96; part II, October, pp. 636–639.

1947

110 "Solution of the General Voltage Regulation Problem by Electrical Analogy," by E. L. Harder, Trans. AIEE, vol. 66, pp. 815–825.

111 "General Analysis of Speed Regulators Under Impact Loads," by G. D. McCann, W. O. Osbon, and H. S. Kirschbaum, Trans. AIEE, vol. 66, pp. 1243–1252.

112 "The Effect of Coulomb Friction on the Performance of Servomechanisms," by G. D. McCann, F. C. Lindvall, and C. H. Wilts, Trans. AIEE, vol. 67, pp. 540–546.

113 "Combination of Calculating Devices as Models With Nonsubstituted Objects as a Method of Investigation of Dynamic Regimes of Complex Systems" (in Russian), by G. L. Polisar, Doklady Akad. Nauk, SSSR 61, No. 4, August, pp. 641–644.

114 "Electronic Servomechanisms Simulator," by F. C. Williams and F. J. U. Ritsan, *Journal of the IEE*, vol. 94, part 2A, pp. 112–124.

1950

115 "A Generalized Analogue Computer for Flight Simulation," by A. C. Hall, Trans. AIEE, vol. 69, part II, pp. 308–320.

116 "An Electronic Simulator for Nonlinear Servomechanisms," by C. M. Edwards and E. C. Johnson, Jr., Trans. AIEE, vol. 69, part I, pp. 300–306.

117 "The Effect of Friction on the Behavior of Servo Mechanisms at Creep Speeds," by J. G. L. Michel and A. Porter, IEE Paper 1033, London, England.

1951

118 "Simulation—Its Place in System Design," by H. H. Goode, Proceedings of the IRE, vol. 39, December, pp. 1501–1506.

119 "Electrical Analogies Applied in the Classification and Investigation of Control Systems" (in Czech), by D. Singer, *Slabopr. Obz.*, vol. 12, pp. 171–178.

1952

120 "Analogue Computers for Feedback Control Systems," by R. A. Bruns, Trans. AIEE, vol. 71, part II, pp. 250–254.

121 "A Differential Analyzer Study of Certain Nonlinearly-Damped Servomechanisms," by R. R. Caldwell and V. C. Rideout, Trans. AIEE, vol. 72, 1953, part II, pp. 165–170.

122 "Mechanical Harmonic Analyzer and Some Applications to Servo Systems," by J. G. Henderson, *Engineering*, vol. 173, January 11–18, pp. 52–53; 68–70.

123 "Analogue-Computer Simulation of an Autopilot Servo System Having Nonlinear Response Characteristics," by A. L. Jones and J. S. White, NACA TN 2707, June.

124 "A Study of Amplifier Saturation and Magnetic Saturation in a Servomechanism," by C. Leondes, AIEE Miscellaneous Paper 52-198, May.

125 "An Electronic Servo Simulator for Unstable and Open Loop Systems," by N. T. VanderWalt, *Electronics Eng.*, vol. 44, February, pp. 52–57.

Nonlinear Systems

1934

126 "Theoretical Foundations of the Automatic Regulation of Temperature," by A. Ivanoff, *Journal of the Institution of Fuel*, London, England, February.

1946

127 "Analysis of Relay Servomechanisms," by H. K. Weiss, *Journal of the Aeronautical Sciences*, vol. 13, July, pp. 364–376.

1947

128 "The Effects of Backlash and of Speed Dependent Friction on the Stability of Closed-Cycle Control Systems," by A. Tustin, *Journal of the IEE*, London, England, vol. 94, part IIA, pp. 143–151.

1948

129 "The Transfer-Locus Method for Servomechanisms With Friction" (in German), by W. Oppelt, *Zeitschrift des Vereines deutscher Ingenieure*, vol. 90, June, pp. 179–183.

1949

130 "Stability of Linear Oscillating Systems With Constant Time Lag," by H. I. Ansoff, *Journal of Applied Mechanics*, Trans. ASME, vol. 71, pp. 158–164.

131 "Methods of Investigating the Stability of Regulating Systems With Distributed Parameters" (in Russian), by V. I. Drozdovich, *Avtomat. i Telemekh.*, vol. 10, pp. 77–83.

132 "An Analysis of Relay Servomechanisms," by D. A. Kahn, Trans. AIEE, vol. 68, part II, pp. 1079–1088.

133 "An Analysis of the Working of Discontinuous Physical Systems"—"and Its Application to Servomechanisms" (in French), by F. H. Raymond, *Annales de Télécommunication*, Paris, France, vol. 4, July, pp. 250–256; August–September, 307–314; October, pp. 347–357.

134 "Theory of Intermittent Regulation," by Ya. Z. Tsipkin, *Avtomat. i Telemekh.*, vol. 10, pp. 189–224.

135 "Instrument Inaccuracies in Feedback Control Systems With Particular Reference to Backlash," by J. Zauderer, H. T. March, and M. Yacther, Trans. AIEE, vol. 68, part I, pp. 778–788.

1950

136 "The Theory of Relay Servomechanisms" (in French), by J. R. Dutilh, *Onde. Elec.*, vol. 30, pp. 438–445.

137 "Stability of Servomechanisms With Linearly Varying Elements," by M. J. Kirby, Trans. AIEE, vol. 69, pp. 1662–1668.

138 "A Frequency Response Method for Analyzing and Synthesizing Contactor Servo-Mechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, pp. 270–283.

139 "Nonlinear Techniques for Improving Servo Performance," by Donald McDonald, Proceedings of the NEC, vol. 6, pp. 400–421.

140 "A New Approach to the Design of Pulse-Monitored Servo Systems," by A. Porter, F. W. Stoneman, and D. F. Lawden, *Journal of the IEE*, London, England, vol. 97, part 2, pp. 597–610.

1951

141 "Extension of the Nyquist Theory to the Case of Nonlinear Characteristics" (in French), by A. Blaguiere, C. R. Acad. Sci., Paris, France, vol. 233, July, pp. 345–347.

142 "Stability of Varying-Element Servomechanisms With Polynomial Coefficients," by M. J. Kirby and R. M. Guilianelli, Trans. AIEE, vol. 70, part II, pp. 1447–1450.

143 "A General Theory of Sampling Servo Systems," by D. F. Lawden, Proceedings of the IEE, London, England, vol. 98, part 4, October, pp. 31–36.

144 "Sampled-Data Control Systems Studied Through Comparison of Sampling With Amplitude Modulation," by W. K. Linvill, Trans. AIEE, vol. 70, part II, pp. 1779–1788.

145 "Backlash and Resilience Within the Closed Loop to Automatic Control Systems," by J. H. Liversidge, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 343–372.

146 "A General Criterion of Stability of Servomechanisms" (in French), by J. Loeb, C. R. Acad. Sci., Paris, France, vol. 233, July, pp. 344–345; also *Annales des Télécom.*, vol. 6, December, pp. 346–356.

147 "Harmonic Analysis of Non-Linear Servo-Mechanisms" (in French), by J. Loeb, C. R. Acad. Sci., Paris, France, vol. 233, October, pp. 733–735.

1952

148 "An Approximate Transfer Function for the Analysis and Design of Pulsed Servos," by R. G. Brown and G. J. Murphy, Trans. AIEE, vol. 71, part II, pp. 435–440.

149 "Equivalences for the Analysis of Circuits With Small Nonlinearities," by W. J. Cunningham, *Journal of Applied Physics*, vol. 23, June, pp. 653–657.

150 "On Some Nonlinear Phenomena in Regulating Systems" (in Russian), by L. C. Goldfarb, *Avtomatika i Telemekhanika*, vol. 8, pp. 349–383; translated in National Bureau of Standards Report 1691, 1952.

151 "Coulomb Friction in Feedback Control Systems," by V. B. Haas, Jr., Trans. AIEE, vol. 72, part II, pp. 119–126.

152 "Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements," by E. Calvin Johnson, Jr., Trans. AIEE, vol. 71, part II, pp. 169–181.

153 "Limiting in Feedback Control Systems," by R. J. Kochenburger, Trans. AIEE, vol. 72, part II, pp. 180–194.

154 "The Use of Nonlinear Feedback to Improve the Transient Response of a Servomechanism," by J. B. Lewis, Trans. AIEE, vol. 71, part II, pp. 449–453.

155 "A General Linearizing Process for Non-Linear Control Systems," by J. M. Loeb, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 275–280.

156 "Non-Linear Control Systems," by N. Minorsky, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., pp. 309–318.

157 "The Analysis of Sampled Data Systems," by J. R. Ragazzini and L. A. Zadeh, Trans. AIEE, vol. 71, part 2, November, pp. 225–234.

158 "Stabilization of a Servomechanism Subject to Large Amplitude Oscillation," by E. S. Sherrard, Trans. AIEE, vol. 71, part 2, pp. 312–324.

1953

159 "Open-Loop Frequency Response Method for Nonlinear Servomechanisms," by R. L. Cosgriff, Trans. AIEE, vol. 72, part II, pp. 222–225.

160 "Describing Function Method of Servomechanism Analysis Applied to Most Commonly Encountered Nonlinearities," by H. D. Grief, Trans. AIEE vol. 72 part II, pp. 243–24

161 "Extension of Conventional Techniques to the Design of Sampled-Data Systems," by W. K. Linvill and R. W. Sittler, Professional Group on Radio Telemetry and Remote Control, IRE Convention, New York, N. Y.

162 "Translation of Papers on Stability of Non-Linear Feedback Control Systems," National Bureau of Standards Report 1691, U. S. Department of Commerce, May 29, 1952.

Books

1944

163 "Automatic Control Engineering," by E. Sinclair Smith, McGraw-Hill Book Company, Inc., New York, N. Y. (Bib.)

1945

164 "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand Company, Inc., New York, N. Y.

165 "Principles of Industrial Process Control," by D. P. Eckman, John Wiley & Sons, Inc., New York, N. Y.

166 "Fundamental Theory of Servomechanisms," by L. A. MacColl, D. Van Nostrand Company, Inc., New York, N. Y.

1947

167 "On the Calculation of Regulating Phenomena in Linear Systems," by Jost Hanny (in German), thesis, Eidgenossische Technische Hochschule, Zurich, Switzerland.

168 "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., Rad. Lab. Series, New York, N. Y., vol. 25.

169 "Servomechanism Fundamentals," by Henri Lauer, Robert Lesnick, and L. E. Matson, McGraw-Hill Book Company, Inc., New York, N. Y.

170 "Principles of Automatic Regulation" (in German), by Winfried Oppelt, Wolfenbutteler Verlagsanstalt, Hannover, Germany.

1948

171 "Principles of Servomechanisms" (Bib.), by G. S. Brown and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y.

172 "Electronic Instruments," Part II, "Instrument Servomechanisms," by I. A. Greenwood, J. Vance Holdam, Jr., and Duncan MacRae, Jr., McGraw-Hill Book Company, Inc., Rad. Lab. Series, New York, N. Y., vol. 21.

173 "Dynamics of Automatic Controls" (Bib.), by R. C. Oldenbourg and H. Sartorius, translated and edited by H. L. Mason, ASME, New York, N. Y.

1949

174 "Servomechanisms" (in French), Tomer I and II, by P. Colombani, G. Lehmann, J. Loeb, A. Plinmellet, and F. H. Raymond, Société d'Edition d'Enseignement Superieur, 99 Boulevard Saint-Michel, Paris 5e, France.

175 "Continuous Regulation Processes" (in German), by Winfried Oppelt, Wissenschaftliche Verlagsanstalt K.G., Hannover, Germany.

1950

176 "Introduction to Servomechanisms," by A. Porter, John Wiley & Sons, Inc., New York, N. Y.

177 "Bibliography of Information on Servo-Mechanisms and Related Subjects," Ministry of Supply, Technical Information Bureau, London, England (TPA3).

1951

178 "Automatic Feedback Control," by W. R. Ahrendt and J. F. Taplin, McGraw-Hill Book Company, Inc., New York, N. Y.

179 "Servomechanisms and Regulating System Design" (Bib.), by Harold Chestnut and R. W. Mayer, John Wiley & Sons, Inc., New York, N. Y., vol. 1.

180 "Fundamentals of Automatic Control," by G. H. Farrington, John Wiley & Sons, Inc., New York, N. Y.

181 "An Introduction to the Theory of Control in Mechanical Engineering," by R. H. MacMillan, Cambridge University Press, Cambridge, England.

182 "Servomechanisms," Selected Government Research Reports, H. M. Stationary Office, Department of Scientific and Industrial Research, London, England V5.

1952

183 "Automatic and Manual Control," by A. Tustin, Academic Press, Inc., New York, N. Y.

1953

184 "Servomechanism Analysis," by G. J. Thaler and R. G. Brown, McGraw-Hill Book Company, Inc., New York, N. Y.

Process Control

185 "Theoretical Foundations of Process Control" (Bib.), by G. H. Farrington, *Journal of the IEE*, London, England, vol. 94, part 2A, 1947, pp. 23–38.

186 "Fundamental Principles of Automatic Regulators and Servomechanisms," by A. L. Whiteley, *Journal of the IEE*, London, England, vol. 94, part 2A, 1947, pp. 5–19.

187 "The Effect of Transmission Distance on the Stability of Flow-Control Processes," by D. B. Kirk, *Instruments*, vol. 23, 1950, pp. 1191–1194.

188 "The Practical Application of Frequency Response Analysis to Automatic Process Control," by C. I. Rutherford, Proceedings of The Institution of Mechanical Engineers, vol. 162, 1950, pp. 334–343.

189 "The Influence of Measuring and Transmission Lags," by A. R. Aikman, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., 1951, pp. 205–216.

190 "Some Recent Developments in Process Control," by W. H. Howe, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., 1951, pp. 217–233.

191 "Practical Application of Servomechanism Techniques to a Process Control Problem," by G. W. McKnight and S. P. Higgins, Jr., ISA Paper No. 51-8-2.

192 "Rule of Thumb Relations for Design and Adjustment of Regulated Processes" (in German), by W. Oppelt, *Chemie-Ing.-Technik*, 1951, pp. 190–193.

193 "The Dynamic Behavior of the Control Stage in Pressure Control" by P. Profos (in German). *Schweiz Arch.*, vol. 17, April, 1951, pp. 114–119.

194 "Stability of Controlled Processes," by J. G. Baring, *Instruments*, vol. 25, April, 1952, pp. 456–457, 494–497.

195 "Frequency-Response Analysis for Industrial Automatic-Control Systems," by D. W. St. Clair, W. F. Coombs, Jr., and W. D. Owens, Trans. ASME, vol. 74, 1952, pp. 1133–1147.

196 "The Development of Automatic Process Control" (Bib.), by A. J. Young, Instruments and Measurements Conference, Stockholm, Sweden, 1952.

Autopilot

197 "Electric Automatic Pilots for Aircraft," by P. Halpert and O. E. Esval, Trans. AIEE, vol. 63, 1944, pp. 861–866.

198 "Frequency-Response Method for Determination of Dynamic Stability Characteristics of Airplanes With Automatic Controls," by Harry Greenberg, NACA Report No. 882, 1947.

199 "Application of Servo Systems to Aircraft," by J. R. Moore, *Aeronautical Engineering Review*, vol. 8, January, 1949, pp. 32–43.

200 "A Method for Predicting the Stability in Roll of Automatically Controlled Aircraft Based on the Experimental Determination of the Characteristics of an Automatic Pilot," by R. T. Jones and L. Sternfield, NACA TN 1901, June, 1949.

201 "Application of Frequency-Response to Aircraft Autopilot Stability" (Bib.), by C. L. Seacord, Jr., *Journal of the Aeronautical Sciences*, vol. 17, August, 1950, pp. 481–495.

202 "The Pulse Method for the Determination of Aircraft Dynamic Performance," by R. C. Seamens, Jr., B. P. Blasingame, and G. C. Clementson, *Journal of the Aeronautical Sciences*, vol. 17, January, 1950, pp. 22–38.

203 "Investigation of Lateral Dynamic Stability in the XB-47 Airplane," by R. J. White, *Journal of the Aeronautical Sciences*, vol. 17, March, 1950, pp. 133–148.

204 "Aerodynamic Stability and Automatic Control" (Bib.), by William Bollay, *Journal of the Aeronautical Sciences*, vol. 18, September, 1951, pp. 569–617.

205 "Design and Analysis of a Conservative Dynamic Load Simulator," by E. H. Gamble and B. W. Hatten, *Journal of Applied Physics*, vol. 22, October, 1951, pp. 1250–1257.

206 "A Survey of Stability Analysis Techniques for Automatically Controlled Aircraft," by A. L. Jones and B. R. Briggs, NACA TN 2275, January, 1951.

207 "Automatic Flight Control—Analysis and Synthesis of Lateral-Control Problem," by R. N. Bretoi, Trans. ASME, vol. 74, 1952, pp. 415–427.

208 "Improving the Dynamic Response of Airplanes by Means of Electric Equipment," by W. R. Monroe, Trans. AIEE, vol. 71, part II, 1952, pp. 441–449.

Test Equipment

209 "Methods of Testing Small Servo Mechanisms and Data Transmission Systems," by E. W. Marchant and A. C. Robb, *Journal of the IEE*, part 2A, vol. 94, 1947, pp. 292–297.

210 "Laboratory Aids for Electro-Mechanical System Development," by G. C. Newton, Jr., and W. T. White, Trans. AIEE, vol. 66, 1947, pp. 315–321.

211 "Instrument to Measure Servomechanism Performance," by H. I. Tarpley, *Review of Scientific Instruments*, vol. 18, January, 1947, pp. 39–43.

212 "Frequency-Response Measurements of a Hydraulic Power Unit," by M. R. Hannah, Trans. ASME, vol. 70, 1948, pp. 525–534.

213 "Modern Servomechanism Testers," by G. A. Korn and T. M. Korn, *Electrical Engineering*, vol. 69, September, 1950, pp. 814–816.

214 "A Phase-Calibrated Variable Frequency Supply for Testing of Servomechanisms," by D. Morris, Proceedings of the IEE, London, England, vol. 97, part 2, 1950, pp. 37–45.

215 "An Amplidyne Phase Shift Oscillator," by J. C. West, *Journal of Scientific Instruments*, vol. 34, October, 1951, pp. 336–339.

216 "A Harmonic-Response-Testing Equipment for Linear Systems," by D. O. Burns and C. W. Cooper, *Journal of the IEE*, London, England, Paper 1386, 1952.

217 "A Pneumatic Sine-Wave Generator for Process Control Study," by D. P. Eckman and J. C. Moise, paper presented at Instrument Society of America annual meeting, September, 1952.

218 "An Automatic Transfer Function Measuring and Recording System," by R. J. Ehret, J. Embree, E. C. Grogan, and E. F. Hochschild, Trans. AIEE, vol. 72, 1953; also available Brown Instrument Division of Minneapolis-Honeywell Regulator Company.

Hydraulic Components and Applications

219 "Contribution to the Theory of Servomotors for Hydraulic Systems" (in Russian), by Z. Blokh, Bulletin, Acad. Sci. USSR Sec. Tech. Sci., vol. 11, January, 1947, pp. 151–158.

220 "Hydraulic Remote Position-Controllers," by J. E. M. Coombes, *Journal of the IEE*, London, England, vol. 94, part 2A, 1947, pp. 270–282.

221 "Hydraulic Variable Speed Transmissions as Servomotors," by G. C. Newton, Jr., *Journal of The Franklin Institute*, vol. 243, June, 1947, pp. 439–469.

222 "On the Automatic Governing of Hydraulic Turbines: Influence of Certain Intervening Characteristics on Stability Conditions," by D. Gaden (in French), Houille Blanche, No. Spec. B, 1949, pp. 717–723.

223 "Comparison of Hydraulic and Electric Servomotors," by G. C. Newton, Jr. Proceedings of the National Conference of Industrial Hydraulics, vol. 3, pp. 64–86, March, 1950.

224 "Electro-Hydraulic Control of Water Turbines," by A. Grade, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., 1951, pp. 503–518.

225 "Analysis of Some Hydraulic Components Used in Regulators and Servomechaniams," by S. Z. Dushkes and S. L. Cahn, Trans. ASME, vol. 74, 1952, pp. 595–601.

226 "Dynamics of Mech'l Feedback-Type Hydraulic Servomotors," by H. Gold, E. W. Otto, and V. L. Ransom, NACA TN 2767, August, 1952.

227 Contributions to Hydraulic Control—1 "Steady-State Axial Forces on Control-Valve Pistons"; 2 "Transient-Flow Forces and Valve Instability," by Shih-Ying Lee and J. F. Blackburn, Trans. ASME, vol. 74, 1952, pp. 1005–1016.

228 "An Analysis of the Dynamics of Hydraulic Servomotors Under Inertia Loads and the Application to Design," by H. Gold, E. W. Otto, and V. L. Ransom, Trans. ASME, vol. 75, 1953, pp. 1383–1390.

229 "Dynamic Operation of a Force-Compensated Hydraulic Throttling Valve," by J. L. Bower and F. B. Tuteur, Trans. ASME, vol. 75, 1953, pp. 583–587.

Application—The Human Operator

230 "Theory of the Human Operator in Control Systems," by K. J. W. Craik, *British Journal of Psychology*, vol. 38, 1947, pp. 56–61, 142–148.

231 "The Nature of the Operator's Response in Manual Control and Its Implication for Controller Design," by A. Tustin, *Journal of the IEE*, London, England, vol. 94, part 2A, 1947, pp. 190–202.

232 "Certain Characteristics of the Human Servo," by F. V. Taylor, *Electrical Engineering*, vol. 68, March, 1948, pp. 235–238.

233 "The Human Transfer Function in Servo Systems," by J. D. North, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., 1951, pp. 473–494.

234 "Some Engineering Aspects of the Mechanism of Body Control," by R. Mayne, *Electrical Engineering*, vol. 70, March, 1951, pp. 207–212.

235 "An Electronic Apparatus for the Study of the Human Operator in a One-Dimensional Closed Loop, Continuous Pursuit Task," by C. E. Warren, P. M. Fitts, and J. R. Clark, Trans. AIEE, vol. 71, part II, 1952, pp. 19–23.

236 "The Human Operator as a Link in Closed-Loop Control Systems," by L. S. Beals, Jr., *Electrical Engineering*, vol. 71, April, 1952, pp. 319–324.

Miscellaneous Applications

Selsyns:

237 "Selsyn Instruments for Position Systems," by T. M. Linvill and J. S. Woodward, *Electrical Engineering*, June, 1934.

238 "The Electric Shaft" (in German), by F. Certel, *Elektrotech. u. Maschinenbau.*, vol. 61, November, 1943, pp. 569–579.

239 "Continuous Control Servomechanisms," by J. T. McNamey, *Electronics*, vol. 17, December, 1944, pp. 118–125.

240 "Selsyn Design and Application," by T. C. Johnson, Trans. AIEE, vol. 64, October, 1945, pp. 703–708.

241 "Principles of Radar" (chapter 12), M.I.T. Radar School Staff, McGraw-Hill Book Company, Inc., New York, N. Y., 1946.

242 "Electrical Accuracy of Selsyn Generator Control Transformer System," by Harold Chestnut, Trans. AIEE, vol. 65, 1946, pp. 570–576.

Motors:

243 "Selecting Electric Servomotors," by R. S. Edwards, *Machine Design*, vol. 21, January, 1949, pp. 104–109.

244 "The Design of D.C. Motors for Use in Automatic Control Systems," by Paul Lebenbaum, Jr., Trans. AIEE, vol. 68, part II, 1949, pp. 1089–1094.

245 "Transient Response of Small 2-Phase Induction Motors," by A. M. Hopkin, Trans. AIEE, vol. 70, part I, 1951, pp. 881–886.

246 "Transfer Function for a 2-Phase Induction Servo Motor," by L. O. Brown, Trans. AIEE, vol. 70, part II, 1951, pp. 1890–1893.

247 "Selection of Motor and Gear Train for Automatic Control Systems," by A. M. Fuchs, *Product Engineering*, vol. 24, March, 1953, pp. 128–134.

Tachometer:

248 "Analysis of the Drag-Cup A-C Tachometer by Means of 2-Phase Symmetrical Components," by R. H. Frazier, Trans. AIEE, vol. 70, part II, 1951, pp. 1894–1906.

Amplidyne:

249 "The Amplidyne Generator, a Dynamoelectric Amplifier for Power Control," by E. F. W. Anderson, M. A. Edwards, and K. K. Bowman, Trans. AIEE, vol. 59, December, 1940, pp. 937–939.

250 "Industrial Applications of Amplidyne Generators," by D. R. Shoults, M. A. Edwards, and E. E. Crever, Trans. AIEE Supplement, vol. 59, December, 1940, pp. 944–949.

251 "Amplidyne Regulating Systems," by B. Adkins, *Journal of the IEE*, London, England, vol. 94, part 2A, 1947, pp. 49–69.

252 "Amplidyne Control of Tension in the Metal Industries," by W. Spence, BTH Activities, vol. 20, 1949, p. 185.

Military:

253 "Gyroscopic Stabilizer for Tank Guns," by C. R. Hanna and L. B. Lynn, *Electrical Engineering*, vol. 63, October, 1944, pp. 355–360.

254 "Some Servo Mechanisms Used by the Royal Navy," by Commander J. O. H. Gardner, *Journal of the IEE*, vol. 94, part 2A, 1947.

255 "Naval Applications of Electrical Remote-Positional Controllers," by N. E. C. Lampert, *Journal of the IEE*, vol. 94, part 2A, 1947, p. 236–251.

Position Systems:

256 "A System Utilizing Coarse and Fine Position Measuring Elements Simultaneously in Remote-Position Control Servo Mechanisms," by J. C. West, Proceedings of the IEE, vol. 99, part 2, April, 1952, pp. 135–147.

257 "Analysis and Construction of a Position Fixing Servomechanism" (in French), by G. Klein, *Ann. Télécom*, vol. 6, November, 1951, pp. 313–324.

Radar:

258 "Electrical Servomechanisms in Radar Equipment" (in Italian), by A. Vallini, *Elettrotecnica*, vol. 38, October, 1951, pp. 456–462.

Demodulator:

259 "High Performance Demodulator for Servomechanisms," by K. E. Schreiner, Proceedings of the National Elec. Conf., vol. 2, 1946; also M.I.T. Electrical Engineering Department Cont., No. 299, 1946.

Textile Industry:

260 "An Industrial Servo-Mechanism," by P. H. Briggs, *Electronic Eng.*, vol. 24, February, 1952, pp. 58–62.

Magnetic Amplifier:

261 "An Improved Magnetic Servo Amplifier," by C. W. Lufey, A. E. Schmid, and P. W. Barnhart, Trans. AIEE, vol. 71, part I, 1952, pp. 281–289.

Engineering Training:

262 "Control Systems," by G. S. Brown and D. P. Campbell, *Scientific American*, vol. 187, September, 1952, pp. 51–64.

Chemical Industry:

263 "Automatic Control in the Chemistry Industry," by J. W. Broadhurst, A. W. Foster, T. C. Brodrich, and G. E. Wheeldon, *Journal of the IEE*, London, England, vol. 94, part 2A, 1947, p. 79.

264 "An Automatic Chemical Plant," by Eugene Ayres, *Scientific American*, vol. 187, September, 1952, pp. 101–115.

Pneumatic Controls:

265 "The Characteristics of Air-Operated Controllers," by A. R. Aikman and C. I. Rutherford, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., 1951, 175–187.

266 "Attenuation of Oscillatory Pressures in Instrument Lines," by A. S. Iberall, *Journal of Research*, National Bureau of Standards, vol. 45, July, 1950, pp. 85–108.

267 "Analysis of Pneumatic Controllers," by J. M. L. Janssen, Cranfield Conference, *Automatic and Manual Control*, Academic Press, New York, N. Y., 1951, pp. 189–200.

Medical-Anesthesia:

268 "The Use of Feedback Systems in the Control of Anesthesia," by R. G. Bickford, *Electrical Engineering*, vol. 70, October, 1951, pp. 852–855.

Automotive-Suspension System:

269 "Control System Engineering Applied to Suspension Systems," by C. J. Martin and R. Jesha, Professional Group on Radio Telemetry and Remote Control, IRE Convention, New York, N. Y., 1953.

Metering:

270 "On the Application of the Feedback Principle in Metering Technique," by E. Lauirla (in Swedish), *Ingenioren*, vol. 60, September, 1951, pp. 598–601.

Frequency Stabilization:

271 "Servo Theory Applied to Frequency Stabilization With Spectral Lines," by W. D. Hushberger and L. E. Norton, *Journal of The Franklin Institute*, vol. 249, May, 1950, pp. 359–366.

Production Systems:

272 "Dynamic Behavior of Linear Production Systems," by D. P. Campbell, *Mechanical Engineering*, vol. 75, April, 1953, pp. 279–283.

Voltage Regulators:

273 "Dynamic Characteristics of Carbon Pile Voltage Regulators," by R. L. Mills, Trans. AIEE, vol. 69, part II, 1950, pp. 1347–1352.

Speed Control:

274 "Wide Speed Range With Servomechanisms," by G. A. Caldwell, *West. Eng.*, vol. 55, September, 1946, p. 157.

275 "Modern Control Dynamics and Stability Criteria as Applied to Gyroscopic Speed Detectors With Hydraulic Drive," by Rene Feiss, Trans. ASME, vol. 66, 1944, pp. 979–988.

276 "Speed Control for the Air-Driven Ultracentrifuge," by Sudhansu D. Majumdar and Gopendra N. Sarkar, *Review of Scientific Instruments*, vol. 19, September, 1948, pp. 541–544.

Fatigue Testers:

277 "Mechanical Oscillators and Their Electrical Synchronization," by S. W. Herwald, R. W. Gemmell, and B. J. Lazon, Trans. ASME, vol. 68, 1946, pp. 713–718.

Automatic Machine Tool:

278 "An Automatic Machine Tool," by W. M. Pease, *Scientific American*, vol. 187, September, 1952, pp. 101–115.

Terminology

279 "Automatic Control Terminology," by D. P. Eckman, *Instrumentation*, vol. 2, 1946, pp. 9–15.

280 "Unified Symbolism for Regulatory Controls," by G. A. Philbrick, Trans. ASME, vol. 69, 1947, pp. 47–67.

281 "Standard Symbols in Control Technique" (in German), by F. Ranke, *Brennstoff-Wärme-Kraft*, vol. 3, November, 1951, pp. 369–373.

282 "Proposed Symbols and Terms for Feedback Control Systems," *Electrical Engineering*, vol. 70, October, 1951, pp. 905–909.

283 "Basic Conceptions of Control and Regulating Techniques and Their Specification" (in German), by G. Vafiadis, *Elektrotech. Z.*, vol. 73, April, 1952, pp. 182–185.

BIBLIOGRAPHY

284 "Bibliography on Feedback Control," Subcommittees on Bibliography on Industrial Control and Feedback Control Systems, Part I—AIEE Paper 53-250; Part II—AIEE Paper 53-251.

Discussion

R. H. MACMILLAN.[5] The author has undertaken a formidable task in attempting a Bibliography of frequency-response literature, which is now so extensive that inevitably it must be highly selective, particularly in regard to foreign publications. His method of classification is convenient and valuable. With regard to the introductory discussion, it may be noted that the "describing function" method employed so effectively by Kochenburger and Johnson to the analysis of nonlinear systems had already been used by Tustin in Britain to study the effects of backlash and speed-dependent friction.[6] This discusser regards it as unfortunate that no publication of the Society of Instrument Technology is listed, as many valuable contributions have appeared in its quarterly Transactions which is one of the foremost periodicals in Britain publishing work of interest to control engineers. Other important omissions are the books by Farrington[7] and Tustin.[8] A valuable introduction to the frequency-response method was the September, 1952, issue of the *Scientific American* devoted exclusively to automatic control and containing articles by many recognized authorities.

JOHN L. BOWER.[9] It is unfortunate that this and similar lists of publications cannot give mention of the work of many scientists who, because of having recorded their contributions only in internal organizational publications or data folders, have received little credit for work that in many cases has been critically important. The work of S. Godet while at the General Electric Company was in the latter category, in my opinion. About June, 1940, or earlier, he wrote a brief paper explaining the application of the Nyquist diagram to the design of stable positioning servomechanisms, and presented applications to specific cases. To my knowledge this was the first writing on that important subject.

It would take further research to determine whether or not his efforts did, in fact, antedate that of Harris, Ferrell, and others named in the paper.

Mr. Godet subsequently explored the field of frequency response of servomechanisms and arrived at a great many of the essential results before publications appeared in any form. While his name cannot appear in the bibliography until his papers are published, it deserves definite mention in the Discussion section of the paper, which attempts a historical summary of the progress of the frequency-response technique.

[5] University Lecturer, University of Cambridge, Cambridge, England.

[6] "Effects of Backlash and Speed Dependent Friction on the Stability of Closed-Cycle Control Systems," by A. Tustin, *Journal of the IEE*, vol. 94, 1947, IIA, pp. 143–151.

[7] "Fundamentals of Automatic Control," by G. W. Farrington, Chapman and Hall, London, England, 1951.

[8] "Direct Current Machines for Control Systems," by A. Tustin, Spon, London, England, 1952.

[9] Engineering Specialist, Research, North American Aviation, Inc., Downey, Calif.

R. M. HUTCHINSON.[10] The author should be commended for limiting the scope of his Bibliography in order to make it more useful for a specific group. It is particularly valuable to have him describe the criterion which he has applied in selecting the material which appears in this Bibliography. It would be even more helpful if he could include a short description of the principal sources tapped by him in his search for references. This would be useful because a researcher using this Bibliography might wish to extend it in a certain direction. In this case it would save time to have knowledge of the ground which had already been covered by the author.

There are two references which appeared after the preparation of the author's manuscript which are of sufficient importance to warrant inclusion in the published version of the paper.

Under Terminology is a very interesting paper by A. J. Young.[11] This includes a comparison of recognized terminology in Britain, the United States, Netherlands, Germany, and Switzerland. It also includes a general discussion of historic development similar to that in the present paper except that it is written from the European viewpoint and describes many European references.

Under Bibliography, part III of reference (284) has now appeared as AIEE Paper 53-393. I believe that this is the final part to the series.

AUTHOR'S CLOSURE

In reply to Mr. John L. Bower, it is indeed unfortunate that a more detailed history of the early development of the frequency-response method is not available. It is interesting to note that Mr. Godet's work was not mentioned by Chestnut and Mayer (179) in their book even though they are in the same company.

For Mr. R. M. Hutchinson's information, the material was gathered primarily from cross-checking the bibliographies in several of the books and all of the articles. *Applied Mechanics Reviews* is the primary source for foreign listings.

Mr. A. J. Young's paper is listed under Process Control (196). It has the notation (Bib.) which means that the article has a noteworthy bibliography.

As Mr. R. H. Macmillan states, Professor Tustin (128) did publish his use of the describing function before Mr. Kochenburger. The discussion at the end of Kochenburger's paper indicates that Tustin's work was not yet known in the United States.

Reference (162) includes a paper by L. C. Goldfarb which employs the describing-function method effectively. This was published originally in 1947 (although not available to Tustin).

The author wishes to thank the discussers for their interest in the Bibliography.

The following references complete the Bibliography through 1953:

BOOKS

1 "Principles of Automatic Control," by F. E. Nixon, Prentice Hall, Inc., New York, N. Y.

2 "Kleines Handbuch Technischer Regelvorgange," by Winfried Oppelt, *Verlag Chemie*, 1954.

NONLINEAR SYSTEMS

3 "Optimization of Nonlinear Control Systems by Means of Nonlinear Feedbacks," by R. S. Nuswander and R. H. MacNeal.

4 "Piecewise Linear Servomechanisms," by J. W. Schwartz, *Applications and Industry*, no. 4, January, 1953, pp. 401–405.

[10] Research Supervisor, Brown Instruments Division of Minneapolis-Honeywell Regulator Company, Philadelphia, Pa. Mem. ASME.

[11] "The Development of Automatic Process Control," by A. J. Young, Trans. Instruments and Measurements Conference, Stockholm, 1952, pp. 428–450.

5 "Analyzing Nonlinear Servos," by J. M. Nightingale, *Machine Design*, vol. 25, August, 1953, pp. 139–142.

6 "The Principle of a Servo-Type Mechanism Requiring Variable Elements," by R. Drenick, Proceedings of the IRE, vol. 41, pp. 373–377.

7 "Stabilization of Nonlinear Feedback Control Systems," by R. L. Cosgriff, Proceedings of the IRE, vol. 41, pp. 382–385.

8 "Some Saturation Phenomena in Servomechanisms With Emphasis on the Tachometer Stabilized System," by E. Levinson, Trans. AIEE, part II, vol. 72, pp. 1–9.

Basic Theory

9 "Relative Stability in Closed-Loop Systems," Trans. AIEE, part II, vol. 72, pp. 22–43.

10 "Reduction of Forced Error in Closed-Loop Systems," by L. H. King, Proceedings of the IRE, vol. 41, 1953, pp. 1037–1042.

Computers

11 "Analysis of Control Systems Involving Digital Computers," by W. K. Linvill and J. M. Salzer, Proceedings of the IRE, vol. 41, pp. 901–906.

12 "Application of Electronic Differential Analyzers to Engineering Problems," by C. A. Meneley and C. D. Morrill, Proceedings of the IRE, vol. 41, pp. 1487–1496.

Pneumatics

13 "Dynamic Behavior of Pneumatic Devices," by L. A Gould and P. E. Smith, Jr., *Instruments*, vol. 26, June, 1953, pp. 886–891; July, 1953, pp. 1026–1033.

Hydraulics

14 "Hydraulic Servos Incorporating a High-Speed Hydraulic-Amplifier Actuated Valve," by R. L. Scrafford, Trans. AIEE, part II, vol. 72, pp. 175–180.

15 "Transient Measurement of Feedback Control Systems," by F. H. Ferguson and C. H. Looney, Trans. AIEE, part II, vol. 72, pp. 110–114.

The Frequency-Response Approach to the Design of a Mechanical Servo

By H. A. HELM,[1] WHIPPANY, N. J.

The frequency-response method of analysis has proved extremely valuable in the design of electronic feedback systems. This paper illustrates the application of this method to the design of completely mechanical systems. Basic relations between the mechanical elements are pointed out and the necessary feedback concepts are developed for mechanical systems. These concepts are illustrated by the actual step-by-step design of an automatic contour-following lathe using entirely mechanical, hydraulic, and pneumatic control elements.

Introduction

MECHANICAL servomechanisms antedate the electromechanical servo of today by many years. The flyball governor invented by James Watt was in common use at a time when only the rudimentary elements of electricity were understood. During the last war the development of radar and electronic computers to control essentially mechanical devices, such as heavy guns and aircraft, took the electronic worker into intimate contact with a field which formerly had been exclusively mechanical. The electronic engineer brought with him into this new (to him) field not only new principles but also new techniques and new methods of analysis. The same type of mathematical equations describes the mechanical (or electromechanical) servo and the feedback amplifier.

From his work with feedback amplifiers[2] the electronic engineer added to the methods of servo design two important tools, the concept of impedance and the methods of frequency-response analysis (1).[3] Rather than present a theoretical treatment of frequency-response analysis as applied to servos, this paper will follow the design development of a typical mechanical servo, illustrating rather than deriving the frequency-response method.

A mechanical servo will contain components which are composed of mass (inertia), compliance (springiness), and friction. These elements will be driven by motive units which are sources of force or torque. By taking each element separately and applying Newton's laws, the relations between applied force and resulting velocity can be derived. In electrical design the relations between current and voltage characterize the behavior of circuits composed of the three basic electrical elements, inductance, capacitance, and resistance. It is logical to assume that force and velocity are the analogs of voltage and current. A comparison of the differential equations relating force and velocity for each of the three basic mechanical elements with those relating voltage and current for the three electrical elements shows that this is indeed the case. In a pneumatic system, pressure and flow are likewise analogs of force and velocity. These relations are summarized in Fig. 1.

From Fig. 1 it is obvious that a force applied to a mass or spring will give rise to a velocity in exactly the same way that a voltage applied across an inductance or capacitance will cause electric current to flow. While it is true that mechanical units will rarely consist of only a single basic element, combinations of elements can be handled readily by using the fact that in a closed system the net force must be zero. In effect, this is a restatement of Newton's third law. One example (as shown in Fig. 1) is that of a force acting on a mass restrained by a spring and damped by friction. Since the total force is equal to the sum of the reaction forces of each of the elements, the differential equation given in Fig. 1 results. Here again the method of attack is exactly the same as in electrical design. Electrically, the sum of the voltage drops across each element must be zero (Kirchhoff's law).

It is apparent that the highly developed methods of electrical network theory can be applied to mechanical systems. A fully detailed development of this is available in the extensive literature on the subject (1, 2). From the point of view of the mechanical engineer, it is unfortunate that most of the literature is concerned with deriving an equivalent electrical network and then carrying out the analysis in electrical terms. This is unnecessarily confusing to the mechanical engineer. Analysis should proceed from the basic mechanical schematic without recourse to artificial conversions. Such a course will be followed throughout this paper.

The derivation of the differential equations which describe a mechanical system can be simplified by introducing the concept of mechanical impedance (8). To illustrate this, consider the example of the series M-f-K system in Fig. 1(j). The differential equation is

$$F(t) = M\frac{d\dot{x}}{dt} + f\dot{x} + K\int \dot{x}dt$$

Here, $F(t)$ is an applied force which is an arbitrary function of time, M is the mass of the system, f is the coefficient of "viscous" friction, and K is the spring constant which is the reciprocal of the compliance C. Further, t is time, x is the displacement of the mass, and $\dot{x}$ is the velocity $(dx)/(dt)$. It should be pointed out here that only the frictional force which is proportional to velocity is assumed to be present. The effect of Coulomb or static friction will be dealt with later. When the methods of operational calculus are applied to this equation, the meaning of impedance is made quite clear. These methods have been made rigorous by the methods of Laplace and Fourier transform analysis (2, 3). For practical applications the relation $d/(dt)$ may be replaced by the operator s and $\int dt$ by $1/s$. Application of these rules to the example of Fig. 1(j) results in

$$F(s) = Ms\dot{x}(s) + f\dot{x}(s) + K\frac{\dot{x}(s)}{s}$$

$$= \left(Ms + f + \frac{K}{s}\right)\dot{x}(s) \quad \text{.....[1]}$$

---

[1] Member, Technical Staff, Bell Telephone Laboratories.

[2] See Patent No. 2102671 by H. S. Black, Bell Telephone Laboratories; also his paper, "Stabilized Feedback Amplifiers," *Bell System Technical Journal*, vol. 13, 1934, pp. 1–18.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division and presented at the Semi-Annual Meeting, Pittsburgh, Pa., June 20–24, 1954, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, April 6, 1954. Paper No. 54—SA-36.

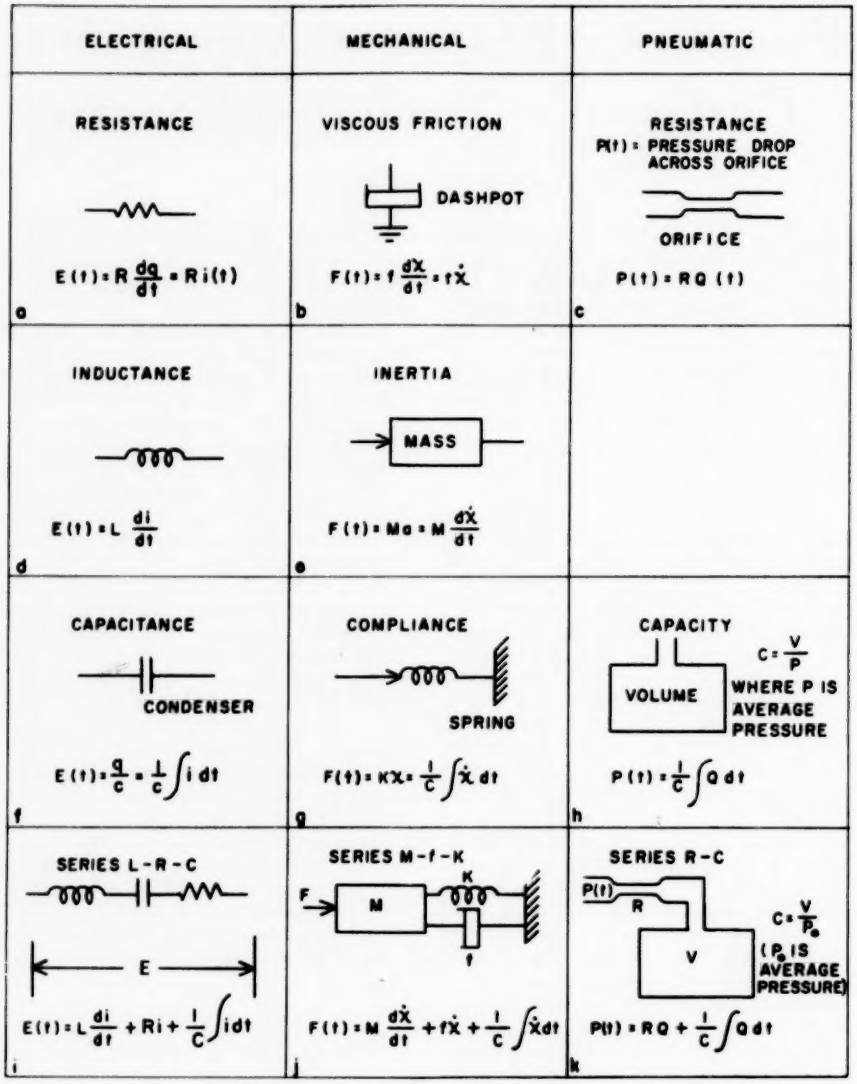


FIG. 1 IMPEDANCES

Here $F(s)$ and $\dot{x}(s)$ are the transforms of $F(t)$ and $\dot{x}(t)$, respectively.

The function $Ms + f + K/s$ clearly relates the applied force and the resulting velocity in operational terms. Equation [1] may be used as a definition of mechanical impedance. In operational form, force equals impedance times velocity. More formally

$$F(s) = Z(s)\,\dot{x}(s) \quad \ldots\ldots\ldots\ldots\ldots\ldots [2]$$

where

$$Z(s) = Ms + f + \frac{K}{s}$$

as in Equation [1].

Returning to the elements of mass, compliance, and friction, the impedance of a pure element of mass is Ms; of friction f; of compliance $1/(sC)$ or K/s The impedance of the elements corresponding to the terms in the parentheses in Equation [1] is thus the sum of the impedance of these elements. The total impedance of any system is determined by the configuration of the system. Mechanical impedances are manipulated in the same manner and follow the same rules as electrical impedances. The foregoing discussion will be adequate for the purposes of this paper and no further elaboration will be given since the subject is well covered in the literature.

FREQUENCY-RESPONSE ANALYSIS

At this point the methods of frequency-response analysis can be introduced. When an arbitrary driving force is inserted in a linear differential equation as in Fig. 1(j), the resulting velocity is the sum of a "transient" term and a "steady-state" term. From Fourier transform theory it can be shown that the steady-state and transient terms of the solution to a linear differential equa-

tion are related. If the behavior of a system when driven by a sinusoidal forcing function is known, the response to almost any forcing function can be derived. This is the method of frequency-response analysis. The solution for a sinusoidal input is relatively easy and once obtained completely describes the system. For this reason frequency-response or steady-state analysis is a very practical method.

The steady-state solution is most readily obtained from the operational form of the differential equations by replacing s with $j\omega$ where $j = \sqrt{-1}$. Thus Equation [2] becomes

$$\left.\begin{aligned} F(j\omega) &= Z(j\omega)\dot{x}(j\omega) \\ Z(j\omega) &= \left(Mj\omega + f + \frac{K}{j\omega}\right) \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [3]$$

Here $F(j\omega)$ and $\dot{x}(j\omega)$ are the complex forms of $F(t)$ and $\dot{x}(t)$ to be used here. By assigning various values to ω and letting $F = A \sin \omega t$, the function $\dot{x}$ can be obtained in vector form. That is, $Z(j\omega)$ "operates" on $\dot{x}$ and changes the magnitude and the angle of $\dot{x}(j\omega)$ to that of $F(j\omega)$.

An example will best illustrate this. By taking the series RC pneumatic circuit of Fig. 1(k) as an example, a result will be obtained which will be useful later (5). Force and pressure are analogous. Let $P(t)$ be the input pressure to the device, and $P_0(t)$ the output pressure, i.e., the pressure in the volume (capacity). Then (see Appendix)

$$P(j\omega) = \left(R + \frac{1}{j\omega C}\right) Q(j\omega)$$

where $P(j\omega)$ and $Q(j\omega)$ are the complex forms of $P(t)$ and rate of flow Q. Since

$$P_0(j\omega) = \frac{Q(j\omega)}{j\omega C} \quad \ldots\ldots\ldots\ldots [4]$$

for the complex form $P_0(j\omega)$ of $P_0(t)$

$$\frac{P_0(j\omega)}{P(j\omega)} = \frac{1}{j\omega RC + 1} \quad \ldots\ldots\ldots\ldots [5]$$

The magnitude of the ratio between output and input is

$$\left|\frac{P_0(j\omega)}{P(j\omega)}\right| = \frac{1}{\sqrt{1 + \omega^2 R^2 C^2}}$$

This quantity is called the "magnitude ratio" of the system ("gain" in the Bell Lab sense). It is not necessarily always greater than unity. In fact, in the foregoing example the magnitude ratio is always unity or less. If there is more than one such network in cascade in the system, the over-all magnitude ratio would be the product of the individual magnitude ratios (assuming no interaction between networks). Since addition is easier than multiplication, it is convenient to express $[P_0(j\omega)]/[P(j\omega)]$ in logarithmic form. We write

$$\frac{P_0(j\omega)}{P(j\omega)} = \left|\frac{P_0(j\omega)}{P(j\omega)}\right| \underline{/ -\theta} \quad \ldots\ldots\ldots\ldots [6]$$

where $\theta = \tan^{-1}(\omega RC)$. Here $\underline{/ -\theta}$ denotes $\cos\theta + j \sin\theta$. One logarithmic form is the decibel (db) which is defined as

$$20 \log_{10} \left|\frac{P_0(j\omega)}{P(j\omega)}\right| = \text{magnitude ratio, expressed in decibels (db)}$$

In Equation [5], for all values of ω such that $\omega RC \ll 1$ (here "$\ll$" means "much less than") the ratio

$$\left|\frac{P_0(j\omega)}{P(j\omega)}\right| \doteq 1, \text{ and } \theta \doteq 0$$

where "$\doteq$" means "is approximately equal to."

For values of ω such that $\omega RC \gg 1$ ("$\gg$" means "much greater than")

$$\frac{P_0(j\omega)}{P(j\omega)} \doteq \frac{1}{j\omega RC}$$

In decibels

$$\left|\frac{P_0(j\omega)}{P(j\omega)}\right|$$

is $-20 \log_{10} \omega RC$. The phase is -90 deg. At the frequency where $\omega RC = 1$ the magnitude ratio is 0.707, i.e., -3db, and the angle is -45 deg.

Taking other points and plotting the results on a scale which is logarithmic versus radian frequency ω on a logarithmic scale, the plot of Fig. 2 is obtained. The time constant τ of the system is RC by definition. It should be noted that the magnitude curve rapidly approaches an asymptote as $\omega\tau$ becomes greater than unity and also approaches zero db axis (magnitude ratio 1) asymptotically as $\omega\tau$ goes to zero. These two asymptotes, if extended, intersect at the point where $\omega\tau$ equals unity as shown by the dotted lines. The phase curve is zero for practical purposes at small values of $\omega\tau$; -45 deg at $\omega\tau = 1$; and -90 deg at the larger values of $\omega\tau$. A great amount of servo design can be accomplished with the manipulation and measurement of simple curves such as these. The curves represent the frequency response of the system of Fig. 1(k) to sine-wave inputs.

Application of Frequency Response to Design

As an example of how the frequency-response methods fit into the actual design and analysis of a servo, the design of an automatic contour-following lathe will be worked through. The basic block diagram of a single-loop servo system is given in Fig. 3 and the fundamental servo equation derived. All the active elements in the servo are grouped together in the block labeled μ. That is, all the frequency-variant elements are included therein, and the block labeled β is assumed in this example to be unity and invariant with frequency. For this case the basic equation becomes (4)

$$\frac{x_0}{x_{in}} = \frac{\mu}{1 + \mu} \quad \ldots\ldots\ldots\ldots [7]$$

A servo system as shown in Fig. 3 must be stable in the following sense. Consider a transient input that dies out. The response of the system must then also die out. Otherwise the system is unstable. It has been shown by the fundamental work of H. S. Black that at least for practical purposes if the magnitude ratio of the open loop is less than unity when the phase shift is -180 deg, no such instability results. Nyquist has expressed in general mathematical terms the explicit conditions for stability in his well-known criterion. In the foregoing argument it is assumed that the output x_0 is subtracted from the input x_{in} when the loop is closed to yield Equation [7].

The concept of stability is a very real and practical matter. Anyone who has seen an unstable servo oscillate or "hunt" obtains a graphic picture of what stability is. In fact, no acceptable servo can be unstable. The servo designer must be concerned not only with the stability of his servo but with the "margin" he has provided against instability, since the components he must use will vary from unit to unit and with age. In addition, he must be concerned not that any transient will eventually decay but how fast

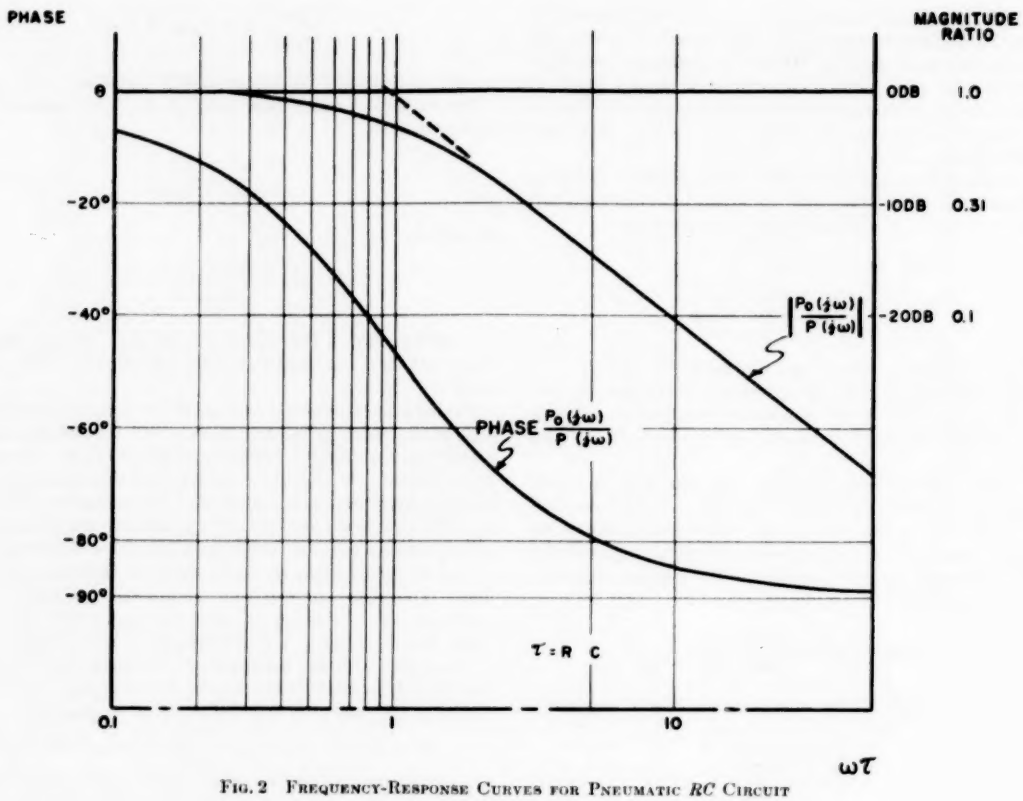
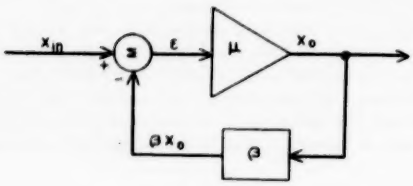


FIG. 2 FREQUENCY-RESPONSE CURVES FOR PNEUMATIC *RC* CIRCUIT



$$\varepsilon = x_{in} - \beta x_o$$

$$x_o = \mu \varepsilon$$

$$x_o = \mu (x_{in} - \beta x_o)$$

$$x_o (1 + \mu\beta) = \mu x_{in}$$

$$\frac{x_o}{x_{in}} = \frac{\mu}{1 + \mu\beta}$$

$$\varepsilon = \frac{x_{in}}{1 + \mu\beta}$$

FIG. 3 FEEDBACK LOOP

it will decay. This matter of "damping" is one of the fundamentals of good servo design. Roughly, a servo with good margins also will be well damped.

To start a servo design, the designer must work to a given set of specifications. In a complex equipment the analysis of the requirements often can be as involved as the actual design. However, in the example chosen, relatively simple specifications are adequate.

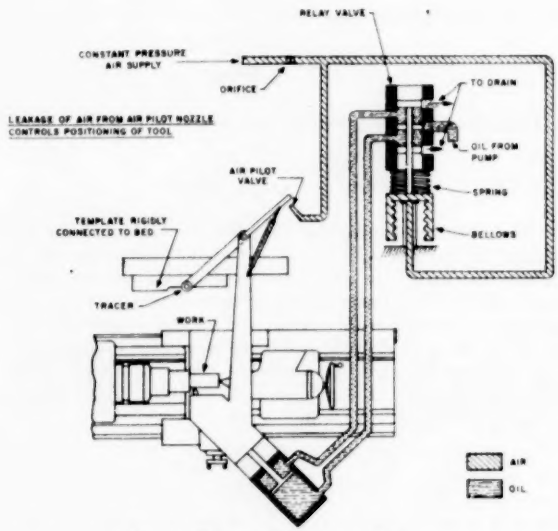


FIG. 4 CONTOUR-FOLLOWING LATHE

Consider the servo shown in Fig. 4. Let it be assumed that the tool carriage of the lathe is to be positioned by a hydraulic piston to agree with a template mounted on the bed of the lathe. The tool must be positioned to an accuracy of ±0.0005 in. under static conditions, that is, with no external back pressure on the tool from the work, and the only loading the internal friction of the

slide and cylinder. The tool slide shall be positioned to an accuracy of ±0.001 in. when the tracing point is moving over the template at a rate of 12 ipm. To eliminate the necessity for case-hardened templates, the tracing point shall not exert a force of more than 3 oz on the template.

Such requirements might possibly be met by the layout shown in Fig. 4. The pneumatic flapper-valve amplifier unloads the tracer point from the output members. There also is no need for flexible hydraulic lines under high pressures to be carried from the moving tracing point to the fixed cylinder. The maximum force which the hydraulic cylinder shall be capable of exerting will be taken to be at least 4 tons (8000 lb).

The appropriate starting point for design is the hydraulic unit. Although at first this might seem like putting the cart before the horse, it is this member which exerts the controlling influence upon the design. The characteristics of the motive element determine the limitations on the servo. If the motive element is underpowered, this limits the job capability at the outset. In addition, some of the characteristics of the motive element affect the errors of the servo in a way which sets the design of the other elements, as will be shown.

The dynamic response of the hydraulic piston and valve is derived in the Appendix. Expressed in operational notation

$$\frac{x_0(s)}{x_v(s)} = \frac{\dfrac{k}{A}}{s(T^2s^2 + 2\zeta Ts + 1)} \quad \ldots\ldots [8]$$

where $x_0(s)$ is the transform of the output position of the piston and $x_v(s)$ is the transform of the displacement of the valve from zero. The constants are given in slightly different forms in the Appendix. The factor

$$(T^2s^2 + 2\zeta Ts + 1)$$

in the denominator is a quadratic function of s. The behavior of the circuit will be determined by the roots of the equation obtained by setting this function equal to zero. If the roots are real and negative the system will be nonoscillatory. If they are complex, the transient response will contain oscillatory terms which will decay with time if the real portions of the roots are negative. The magnitude of the real part of a complex root determines the rate of decay of the oscillatory transient. In Equation [8] the factor T is the frequency-determining element; ζ determines the damping of the system. If $\zeta \geq 1$, the system will be nonoscillatory. For the constants chosen in this paper, ζ is 0.25 which makes the transient response of the system a well-damped oscillation.

For the moment, let us assume that $T \ll 1$ so that Equation [8] may be approximated by

$$\frac{x_0(s)}{x_v(s)} = \frac{\dfrac{k}{A}}{j\omega} \quad \ldots\ldots [9]$$

$s = j\omega$ for the steady state

Making a further assumption that the other elements in the loop contribute only amplification which is frequency invariant, then $x_v(s) = k_2\epsilon$ where ϵ is the error signal as in Fig. 3. Substitution of these approximations in Equation [7] and setting $\omega_0 = k_2(k/A)$

$$\frac{x_0}{x_i} \doteq \frac{\dfrac{\omega_0}{j\omega}}{1 + \dfrac{\omega_0}{j\omega}} \quad \ldots\ldots [10]$$

Fig. 5 is a logarithmic plot of $\mu = \omega_0/(j\omega)$. At $\omega = \omega_0$, we have $|\mu| = 1$ and the curve crosses the zero db axis (magnitude ratio 1). The phase is constant at —90 deg. Previously it was pointed out that multiplying by $1/s$ or $1/(j\omega)$ corresponds to an integration. Since opening the valve causes the piston to move with a velocity proportional to valve displacement and the servo compares an output position with an input position, there obviously must be an integration involved. This integration plays a very important part in the errors of the servo. It can be readily shown that the error ϵ of the servo is

$$\epsilon = \frac{x_i}{1 + \mu} \quad \ldots\ldots [11]$$

and for this case

$$\epsilon(j\omega) = \frac{x_i}{1 + \dfrac{\omega_0}{j\omega}} \quad \ldots\ldots [12]$$

where $\epsilon(j\omega)$ is the complex form of ϵ.

The error for the case of a constant signal and no external forces can be obtained by letting ω approach zero in Equation [12]. Under these conditions the error is zero.

The meaning of ω_0 can be made clearer if it is thought of as the amplification of the system. That is, if an error signal ϵ exists at the input to the amplifier, the motor element will move with a velocity $\omega_0\epsilon$. While the motor element can stand still with no error, to run it will require an error. Thus the motor will lag behind a signal which is changing at a constant rate until $\omega_0\epsilon$ equals the rate of change of the signal. This defines ω_0 as the velocity constant which more formally is

$$\omega_0 = \frac{\text{velocity}}{\text{lag error}}$$

If the loop gain ω_0/ω is plotted logarithmically, the result will be a straight line sloping downward at a rate of 6 db/octave (with a —1 slope as shown) and crossing the zero db axis (the line for which the magnitude ratio is 1) at a radian frequency of ω_0 radians/sec. The phase is constant at —90 deg.

The simplification of μ which was made in Equation [12] is not adequate for much more than illustration. However, if μ is assumed to have the form

$$\frac{\omega_0\omega_i}{j\omega(j\omega + \omega_1)} \quad \text{with} \quad \omega_0 >> \omega_1$$

a reasonably good approximation can be made for a large class of servos. The error equation then becomes

$$\epsilon(j\omega) = \frac{x_i(j\omega)}{1 + \dfrac{\omega_0\omega_1}{j\omega(j\omega + \omega_1)}} \quad \ldots\ldots [13]$$

Under the foregoing assumptions the static error is still zero. The dynamic errors can be evaluated by means of the so-called "error coefficients." It has been shown (4) that Equation [13] may be expanded in a power series in terms of powers of $j\omega$

$$\epsilon(j\omega) = \frac{j\omega}{\omega_0} x_i(j\omega) + \frac{(j\omega)^2 x_i(j\omega)}{\omega_0\omega_1} - \frac{2(j\omega)^3 x_i(j\omega)}{\omega_0^2\omega_1} \quad \ldots [14]$$

Since $(j\omega)^n$ is equivalent to $d^n/(dt^n)$ in the steady state with which we are concerned

$$\epsilon(t) = \frac{\dot{x}_i(t)}{\omega_0} + \frac{\ddot{x}_i(t)}{\omega_0\omega_1} -$$

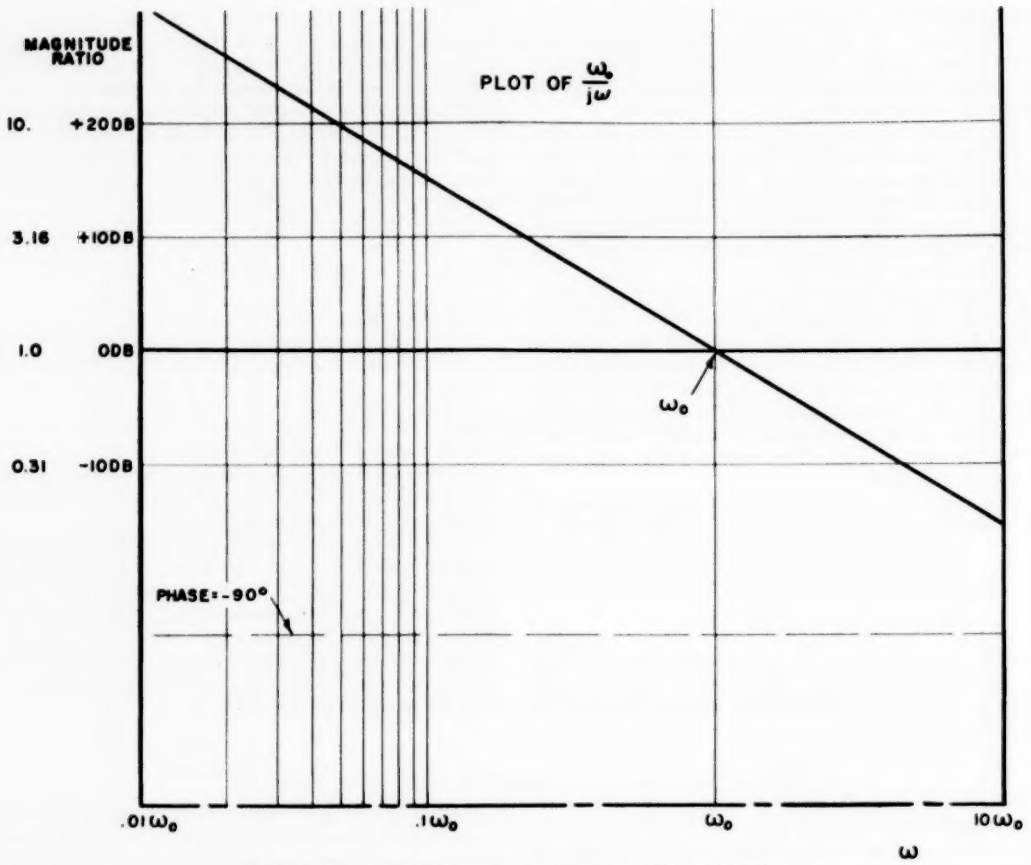


FIG. 5 APPROXIMATE FREQUENCY RESPONSE OF A HYDRAULIC UNIT

This shows that the principal errors are first, a "velocity" lag; second, an "acceleration" lag. In the specifications there is no requirement on acceleration, but there is an implied velocity requirement. The requirement that error shall be no greater than 0.001 in. when the template causes the input tracer to move at 12 ipm implies

$$\omega_0 = \frac{\frac{12}{60}\text{ips}}{0.001\text{ in.}} = 200 \text{ radians/sec}$$

The response of the hydraulic piston can now be evaluated. We assume that the mass of the tool slide and piston is 64 lb (2.0 slugs) the diameter of the piston is 5.0 in., and the total volume of oil under compression is 0.0064 ft³. Equation [8] is evaluated for these constants in the Appendix as

$$\frac{x_0(j\omega)}{x_i(j\omega)} = \frac{\frac{k}{A}}{j\omega[2.78 \times 10^{-6}(j\omega)^2 + 8.36 \times 10^{-4}j\omega + 1]} \quad \ldots\ldots [15]$$

The logarithmic plot of the function

$$\frac{x_0(j\omega)}{x_i(j\omega)} = j\omega \frac{x_0(j\omega)}{x_i(j\omega)}$$

is given in Fig. 6 for $k/A = 1$. It is apparent that the hydraulic system is a very wide-band system. From this it can be concluded that the hydraulic system will not be the limiting factor in the dynamics of the servo. With the piston area of 19.6 sq in. and an assumed pressure of 1000 psi pump pressure, the piston is capable of exerting 19,600 lb force which is a margin of better than two over requirements. The frequency response of the pneumatic section must now be calculated.

SYSTEM WITH PNEUMATIC COMPONENTS

A typical nozzle-flapper pneumatic amplifier is shown in Fig. 7. It is similar to the pneumatic series R-C circuit discussed previously. In fact, the amplifier may be reduced to the same form by deriving an equivalent series resistance R_s (pneumatic) and a frequency-invariant gain μ_p. The formulas for R_s and μ_p are derived in the Appendix

$$\mu_p = \frac{6P_0}{D_n} \quad \ldots\ldots [16]$$

$$R_s = \frac{(P_0)^{1/3}}{1500 D_s^2}$$

where P_0 is the "average" pressure in the bellows. Here, P_0 is one half of the supply pressure and D_s and D_n are the diameters of the orifice and flapper nozzle, respectively. Picking the value of 0.04 in. for D_n and 0.028 for D_s and a supply pressure of 60 psig, the value of μ_p is 5110 and R_s is 1.46 pneumatic ohms. To com-

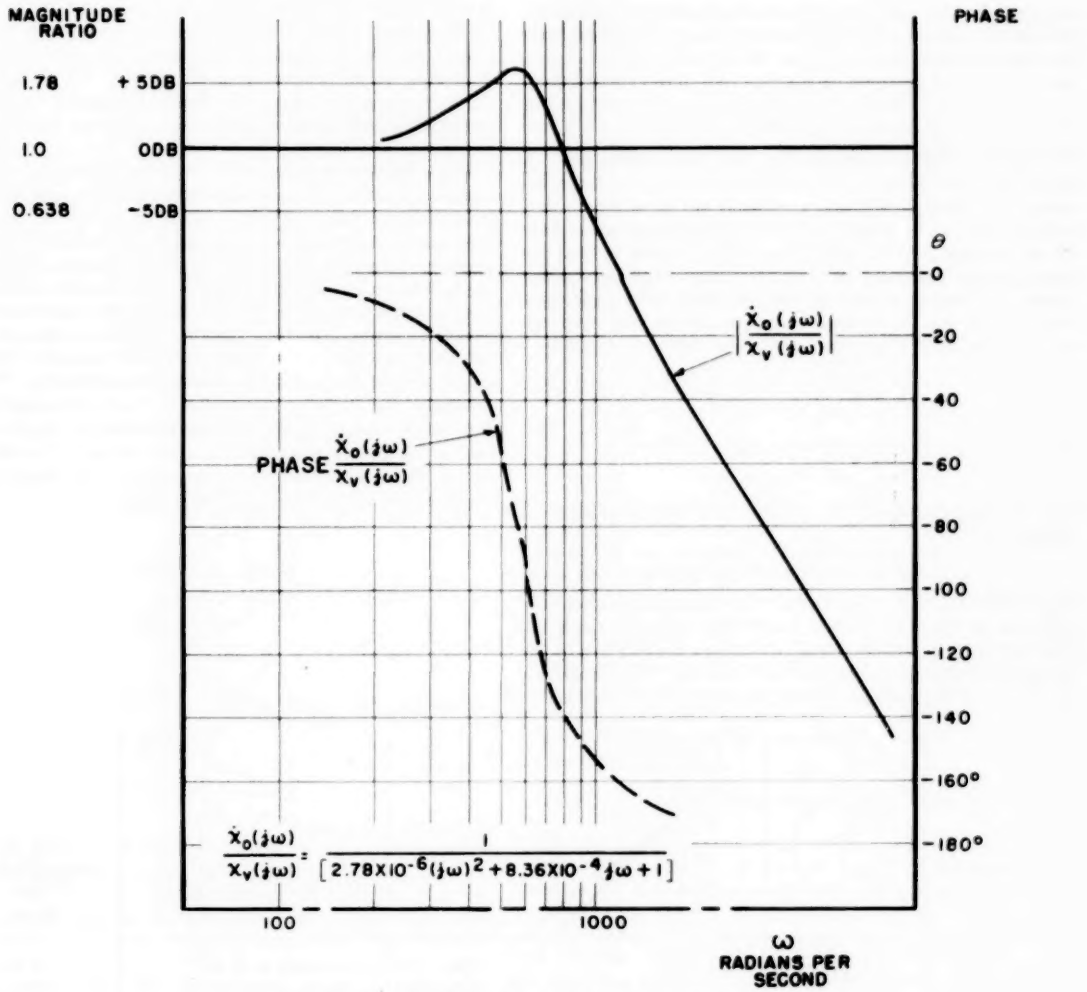


FIG. 6 FREQUENCY RESPONSE OF A HYDRAULIC UNIT

plete the analysis of the pneumatic amplifier, the characteristics of the bellows and valve assembly must be taken into account. The following constants are assumed:

Effective diameter of bellows = 1.0 in.
Effective area of bellows, A_β = 0.79 sq in.
Effective volume of bellows = 0.4 cu in.
Supply pressure, P_s = 60 psi
Spring constant, K_s = 40 lb/in.

The constants of the amplifier were so picked that at null balance the output pressure is equal to $P_s/2$ or 30 psi. The foregoing value of the spring constant includes the reaction force of the valve. With the spring constant and bellows area as stated, the bellows will exert a force of 24 lb upon the spring, compressing it 0.6 in. The action of the spring which loads the bellows increases the pneumatic capacity of the bellows. Let V_1 be the mass (in properly chosen units) of the air in the bellows when the absolute pressure is P_0 + 15 psi. The formula for pneumatic capacity when the bellows is spring-loaded is (see Appendix)

$$C_a = \left[\frac{V_1}{(P_0 + 15)} + \frac{A_\beta^2}{K_s}\right] \quad [17]$$

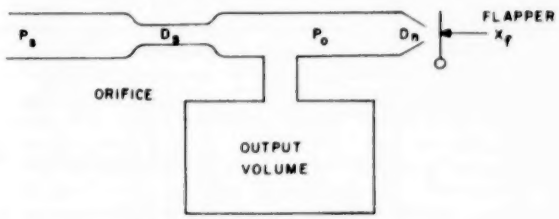


FIG. 7 NOZZLE-FLAPPER PNEUMATIC AMPLIFIER

The output volume in Fig. 7 is actually not fixed. This volume is that of the bellows and an average value is used. The spring loading the bellows is not shown in the figure. Making use of Equation [5], the response (7) of the pneumatic amplifier to a change in the flapper position can be written (see Appendix)

$$\frac{P_0(j\omega)}{x_f(j\omega)} = \frac{\mu_p}{j\omega R_s C_p + 1} \quad [18]$$

Here $x_f(j\omega)$ is the complex form of the flapper co-ordinate x_f. Note that x_f is the x_{in} of Fig. 3. Substitution of the given values into Equations [17] and [18] gives

$$\frac{P_0(j\omega)}{x_f(j\omega)} = \frac{5110}{j\omega(0.0359) + 1} \quad \ldots\ldots [19]$$

The radian frequency at which $\omega R_a C_p = 1$ is 28 radians/sec. A logarithmic plot of the over-all response μ can now be made, assuming that the hydraulic-valve constant can be so chosen that sufficient gain can be obtained to make the magnitude curve cross the zero db-axis (line for which the magnitude ratio is 1) at the design objective of $\omega_0 = 200$ with the slope -1 (6 db/octave). To test whether or not the pneumatic amplifier will be a limiting factor, the hydraulic section can be assumed to vary with frequency as $\omega_0/(j\omega)$. The response function plotted in Fig. 8 is

$$\left.\begin{aligned} \mu &= \left(\frac{\omega_0}{j\omega}\right)\left(\frac{28}{j\omega + 28}\right) \\ &= \left(\frac{200}{j\omega}\right)\left(\frac{28}{j\omega + 28}\right) \end{aligned}\right\} \quad \ldots\ldots [20]$$

Here we have multiplied the transfer functions of elements in cascade, i.e.

$$[P_0(j\omega)/x_f(j\omega)] \text{ by } [x_0(j\omega)/P_0(j\omega)]$$

Note that x_0 is proportional to P_0. The dotted extension of the original slope intersects the axis at $\omega_0 = 200$. As can be seen, the solid curve, which is the function $|\mu|$, crosses the axis at $\omega = 80$. At this point there is a phase shift of -162 deg. The phase margin is only 18 deg (it should be at least 30 deg; see reference 9). If there were only -18 deg of phase shift contributed by the rest of the system, the servo would become unstable. The hydraulic system, although very good, will certainly contribute some phase shift even at the frequency of 80 rad/sec. In addition, any change in parameters due to production tolerances, variations in pneumatic supply pressure, and so on, will again tend to reduce the margin of stability. At this point it must be realized that with the simple system proposed, the design objectives cannot be realized. The choice then is between reducing the requirements or improving the design.

Since it is the bellows assembly that loads down the pneumatic amplifier, a logical approach would be to isolate the amplifier from the bellows by a "pilot relay" between the amplifier and bellows. Such a device is shown in Fig. 9, together with a derivation of its response equation (Fig. 13). This device operates upon the "null-balance" principle. In effect, it is a small servo in itself. Its output pressure equals its input pressure very closely. The diaphragm of Fig. 9 plays a role similar to that of the flapper of Fig. 4, in that it restricts the flow through the nozzle of Fig. 9 so that x in Fig. 9 is like the co-ordinate x_f of the flapper. The first equation in Fig. 9 is derived in the same manner as Equation [18]. The response of the pilot relay is

$$\frac{P_x(j\omega)}{P_0(j\omega)} = \frac{\dfrac{\mu_0 A}{K}\,\tau_1}{j\omega\tau_1 + 1} \quad \ldots\ldots [21]$$

$$\tau_1 = \frac{RC}{1 + \dfrac{\mu_0 A}{K}}$$

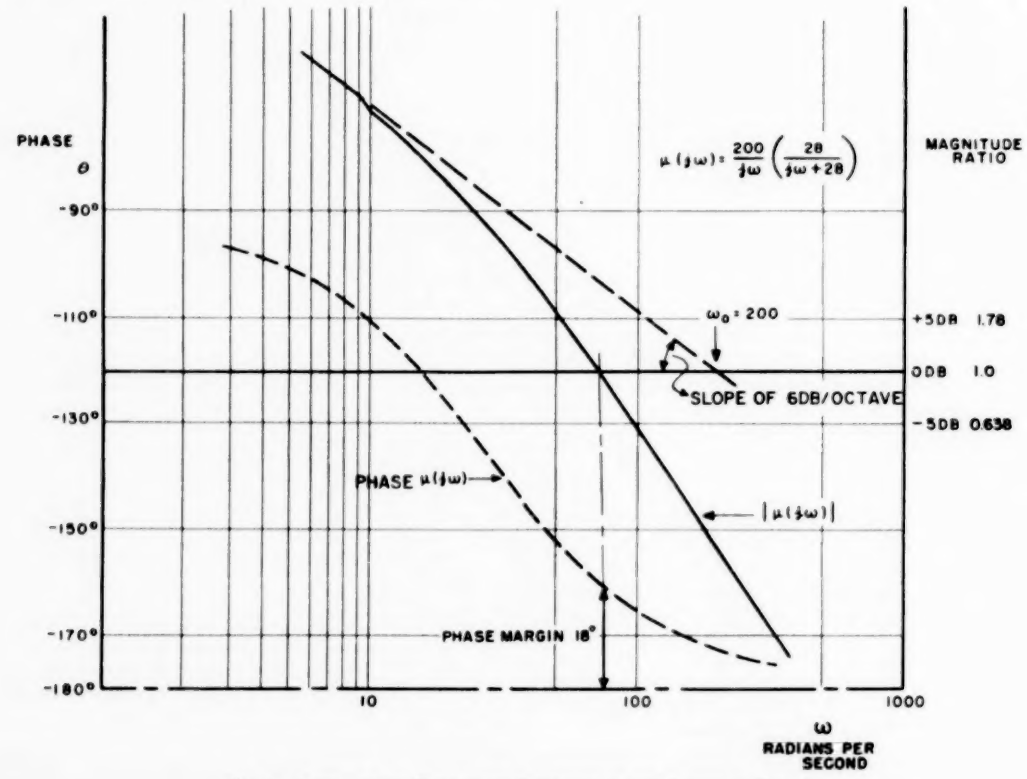
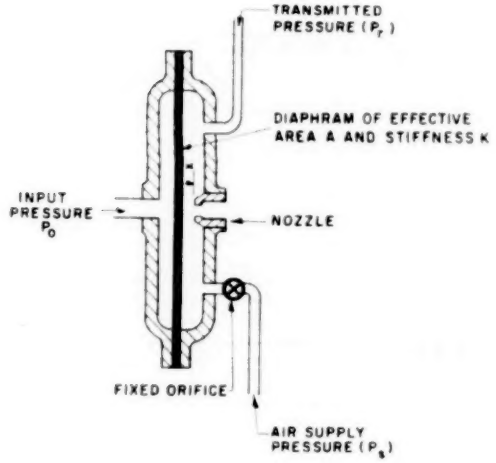


Fig. 8 Approximate Frequency Response of the Lathe of Fig. 4



$$P_r = \frac{\mu_0 x}{sCR+1} = \frac{\mu_0 x}{\tau s+1} \qquad \tau = RC$$

$$x = \frac{A}{K}(P_0 - P_r)$$

$$P_r = \frac{\mu_0 A/K}{\tau s+1}(P_0 - P_r)$$

$$\frac{P_r}{P_0} = \frac{\frac{\mu_0 A/K}{\tau s+1}}{1+\frac{\mu_0 A/K}{\tau s+1}} = \frac{\mu_0 A}{K\tau}\,\frac{1}{s+\left(\frac{1+\mu_0 A/K}{\tau}\right)}$$

FIG. 9 PILOT RELAY

where P_0 is the input pressure to the relay and P_r is the pressure transmitted to the bellows. Here μ_0 is the amplification factor of the pilot relay, R is the output resistance of the relay, C is the volume capacity loading the relay (due to the bellows and spring in Fig. 13), A is the area of the diaphragm, and K is the spring constant of the diaphragm. If the same characteristics of supply pressure, nozzle, and orifice diameters are chosen as in the previous amplifier, the time constant τ is multiplied by the factor

$$\frac{1}{1+\frac{\mu_0 A}{K}}$$

to obtain the new time constant τ_1. Therefore the time constant of the system is effectively reduced by the factor

$$\frac{1}{1+\frac{\mu_0 A}{K}}$$

Since the pilot relay will form the load on the flapper amplifier, it is advantageous to make the diaphragm in the pilot relay very stiff. Assuming an effective diameter of the diaphragm as 1.0 in. and a stiffness of 250 lb/in., the factor $(1 + \mu_0 A/K)$ is 17.2. The frequency at which $\omega\tau_1 = 1$ is now 28×17.2, i.e., 428 rad/sec. This value is now comparable in response time to the hydraulic section. We now have for practical purposes

$$\frac{P_r(j\omega)}{P_0(j\omega)} = \frac{482}{j\omega + 482}$$

The loading of the flapper amplifier by the pilot relay can now be calculated. Since the diaphragm moves only a few thousandths of an inch, the volume can be made very small and the pilot relay located very close to the amplifier to eliminate the added volume of pneumatic lines. Using Equation [18] and recalculating $R_s C_s$ gives $R_s C_p = 0.0069$, clearly an improvement in the amplifier time constant. The frequency at which $\omega R_s C_p = 1$ is now 145 rad/sec. Thus

$$\frac{P_0(j\omega)}{x_f(j\omega)} = \frac{5110 \times 145}{j\omega + 145}$$

Taking into account the hydraulic system (since the time constants are now comparable with the constants given in the Appendix), the loop response is given by

$$\mu(j\omega) = \left(\frac{200}{j\omega}\right)\left(\frac{145}{j\omega+145}\right)\left(\frac{482}{j\omega+482}\right)\left[\frac{1}{2.78\times10^{-6}(j\omega)^2 + 8.36\times10^{-4}j\omega + 1}\right] \quad [22]$$

This function is plotted in Fig. 10. As can be seen, at the point of gain crossover there is a phase shift of —159 deg giving a phase margin of 21 deg as shown. This includes the response of all elements which the function of Fig. 8 did not. At the frequency at which the phase shift becomes —180 deg, the magnitude ratio is 0.6, i.e., —5 db approximately, whence the gain margin is 1.6 (5 db approximately). The gain margin should be at least 2.5 (8 db) (see reference 9). If the gain were to increase by 5 db or a little less than 2:1 the servo would become unstable. This is clearly not a practical design as yet.

The right number of Equation [22] was obtained by multiplying the transfer functions

$$[P_0(j\omega)/x_f(j\omega)],\ [P_r(j\omega)/P_0(j\omega)],\ \text{and}\ [x_0(j\omega)/P_r(j\omega)]$$

It might at first appear that the addition of the pilot relay was a waste of time. This is not the case, as a very simple pneumatic network can now be inserted between the flapper amplifier and the pilot relay. In Fig. 11 the device is shown and its response function derived. The amplitude plot of its response is shown in Fig. 12. The effect of such a network is to narrow the band of the servo in a region where the phase shift is not critical. At the higher frequencies the resistance R_2 is much greater than $1/(\omega C)$ and the effect is as if the unit were open to the atmosphere.

In other words, the output pressure is dependent upon the pressure drop across R_2 only. At extremely low pressures the volume C is completely filled—there is no pressure drop through R_1. The derivation given in Fig. 11, relating pressures P_1 and P_2, assumes that there is no capacity loading upon the output pressure. Of course the input capacity of the pilot relay is directly shunting R_2 and C, Fig. 13. However, the capacity of the relay is very small compared to the values to be used for R_2 and C, and therefore its effect may be neglected. The output of the flapper amplifier already has been shown to have a resistance of 1.46 pneumatic ohms. This resistance can be used in place of R_1. The constants selected for this network are

$$R_s = R_1 = 1.46$$
$$C = 0.595$$
$$R_2 = 0.21$$

and

$$\frac{P_2(j\omega)}{P_1(j\omega)} = \left(\frac{j\omega+8}{8}\right)\left(\frac{1}{j\omega+1}\right) \quad [23]$$

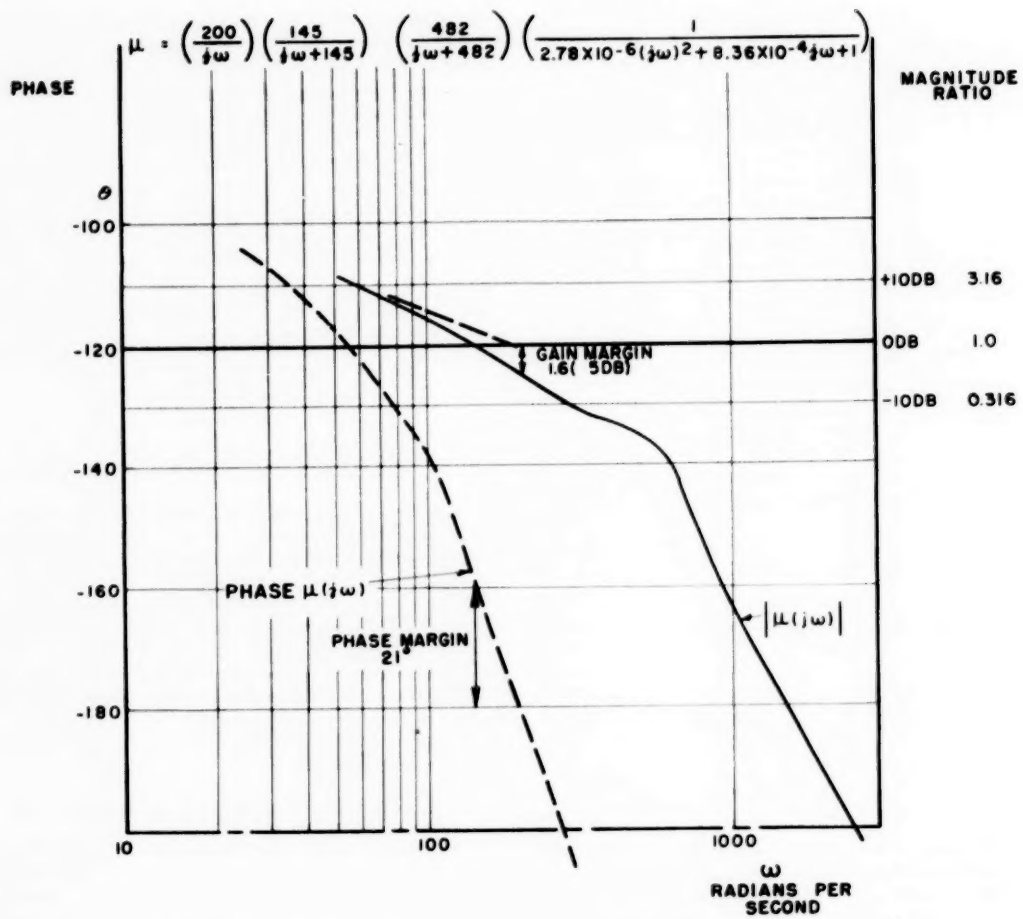
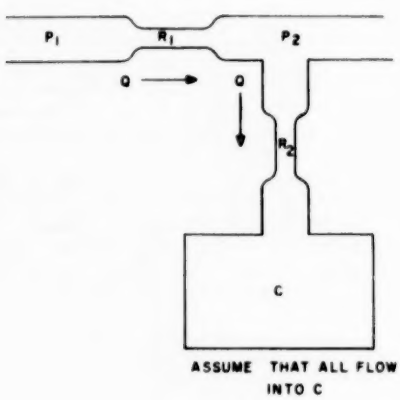


FIG. 10 FREQUENCY RESPONSE OF THE LATHE OF FIG. 4 WITH A PILOT RELAY



$$R_1 Q = P_1 - P_2$$

$$P_2 = R_2 Q + \frac{1}{C}\int Q\,dt$$

$$= R_2 Q + \frac{Q}{sC}$$

$$Q = \frac{P_2}{R_2 + \frac{1}{sC}}$$

$$R_1 Q = \frac{R_1 P_2}{R_2 + \frac{1}{sC}} = P_1 - P_2$$

$$R_1 P_2 = P_1\left(R_2 + \frac{1}{sC}\right) - P_2\left(R_2 + \frac{1}{sC}\right)$$

$$P_2\left(R_1 + R_2 + \frac{1}{sC}\right) = P_1\left(R_2 + \frac{1}{sC}\right)$$

$$\frac{P_2}{P_1} = \frac{sCR_2 + 1}{sC(R_1 + R_2) + 1}$$

FIG. 11 RESISTANCE-CAPACITY CIRCUIT IN LATHE

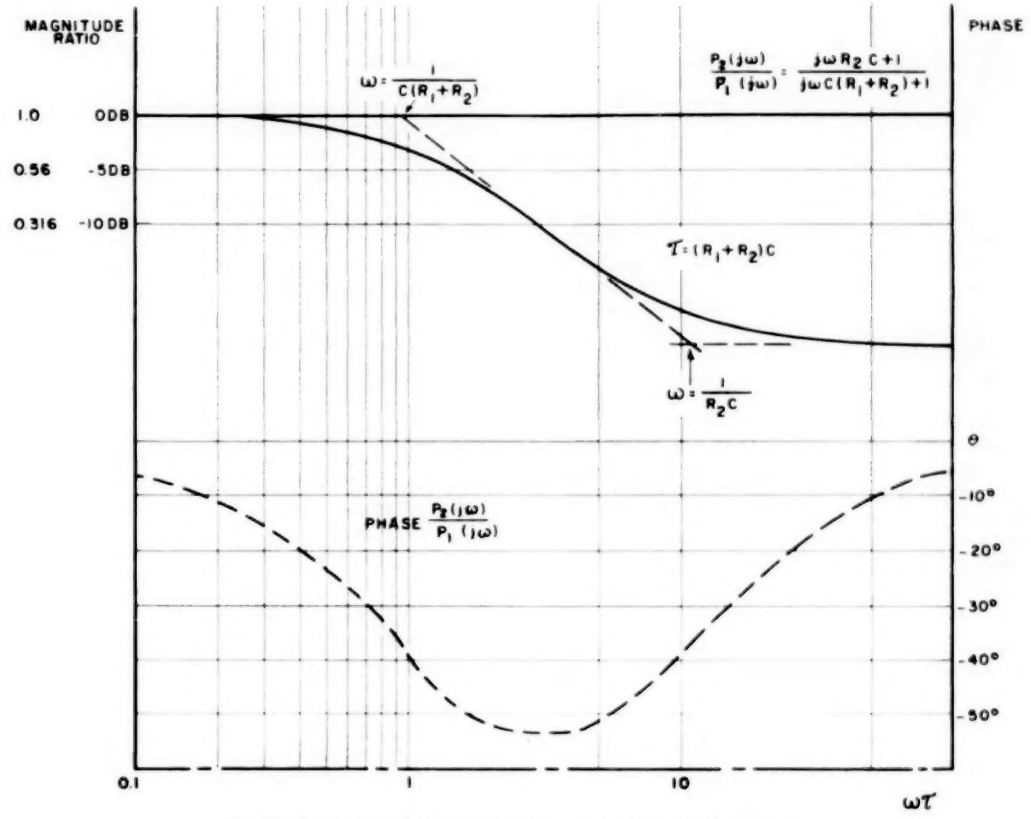


Fig. 12 Frequency-Response Curves for the Circuit of Fig. 11

The logarithm of this function is shown in Fig. 12. (See the complete mechanical schematic of the servo given in Fig. 13, which shows both the pilot relay and the network in their respective places in the loop.)

As pointed out, the input capacity of the pilot relay is small in comparison with the added capacitance, C. Therefore the added network can be considered to a good approximation to form the entire load on the flapper amplifier. This will change the transform

$$\frac{P_0(j\omega)}{x_f(j\omega)}$$

In Equation [18] the frequency response of the flapper amplifier was derived by considering the effect of the load capacitance upon the internal resistance R_i. It has been pointed out already that the resistance R_i can be used as R_1 of Fig. 11. In this case the load on the flapper amplifier is composed of the combination of R_1 and C. We can identify $\mu_f x_f$ in Equation [18] with the "driving" $P(t)$ of Fig. 1(*k*). In a similar manner, for the case where the internal resistance of the amplifier is used as R_1 of Fig. 11, the quantity P_1 can be identified with $\mu_f x_f$ (see Appendix). From Fig. 13 the quantity P_2 is the input pressure to the pilot relay and thus can be considered to be equivalent to P_0 of the earlier argument. The transfer function of the flapper amplifier (flapper to pilot relay) then becomes

$$\frac{P_0(j\omega)}{x_f(j\omega)} = \frac{5110}{(j\omega+1)} \frac{(j\omega+8)}{8}$$

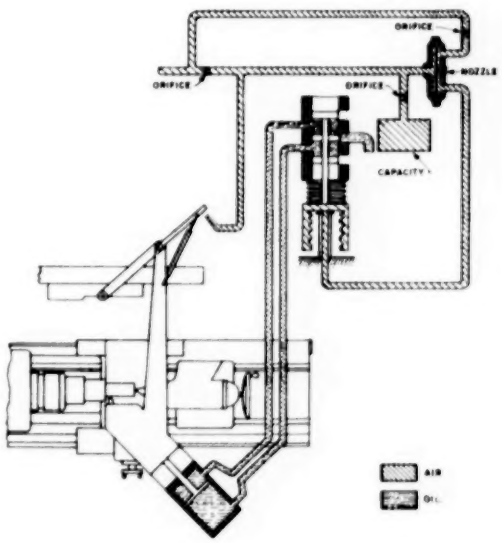


Fig. 13 Final Lathe Design

and the over-all loop transfer function is now

$$\mu(j\omega) = \left(\frac{200}{j\omega}\right)\left(\frac{1}{j\omega + 1}\right)\left(\frac{j\omega + 8}{8}\right)\left(\frac{482}{j\omega + 482}\right)\left[\frac{1}{2.78 \times 10^{-6}(j\omega)^2 + 8.36 \times 10^{-4}j\omega + 1}\right]$$

This equation looks formidable indeed, but actually it is quite simple to evaluate graphically as the logarithms of the magnitudes add as do the angles. Each factor is drawn by a standard curve and all are added together to obtain the final plot in Fig. 14.

At the point of unity magnitude ratio (0 db) the phase shift is —119 deg, giving a phase margin of 66 deg. With this amount of margin the system can stand a considerable amount of variation without becoming unstable. At the frequency at which the phase shift is —180 deg the magnitude ratio is 0.05, i.e., —26 db. The gain margin is 1/0.05, i.e., 20, or 26 db. The gain may be increased by 20 times without causing "hunting" or oscillations. More important, with this margin the servo will be well clamped and recover from transients with a minimum number of overshoots. To complete the design, only the constants must be evaluated to give the required ω_0 of 200 rad/sec.

A 0.001-in. steady displacement will cause an increase in pressure of (set $\omega = 0$ in Equation [19])

$$0.001 \times \mu_s = 5.11 \text{ lb}$$

Since there is no loss in the network or in the pilot relay for steady pressures, this will result in a pressure also of 5.11 lb in the valve bellows. With an area of 0.79 sq in., there is a force of approximately 4 lb on the bellows spring compressing it by 0.1 in. This, then, is a gain of 100:1 from flapper position (null point) to hydraulic valve. An ω_0 of 200 requires that a displacement of the null point of 0.001 in. cause the output member to move at a velocity of 0.2 ips. The flow of oil required in cubic inches per second is equal to $A\dot{x}_0$ where A is the area of the piston. For the foregoing rate and the piston area of 19.6 sq in. assumed previously, a flow of 3.92 cu in. per sec is required. This will occur when the valve is opened 0.1 in., giving a valve constant of 39.2 $in^3/sec/in.$

This completes the design, except for the evaluation of the static errors. At first glance this might appear to be rather late in the design to determine a basic requirement, but actually the static error depends upon the final loop gain and upon the static friction upon the output member. The latter, of course, can only be evaluated accurately after construction of the cylinder and piston.

The transfer function $\mu(j\omega)$ in the foregoing is obtained by multiplying the transfer functions $[P_0(j\omega)/x_f(j\omega)]$, $[P_c(j\omega)/P_0(j\omega)]$, and $[x_0(j\omega)/P_c(j\omega)]$ as before.

The static error Δ_T is

$$\Delta_T = \frac{F_s}{j\omega Z(j\omega)} \frac{1}{1 + \mu(j\omega)} \quad [24]$$

F_s is the force necessary to break the static friction; $Z(j\omega)$ is the mechanical impedance of the output member (without feedback) as defined in Equation [2]. Since the impedance relates the force

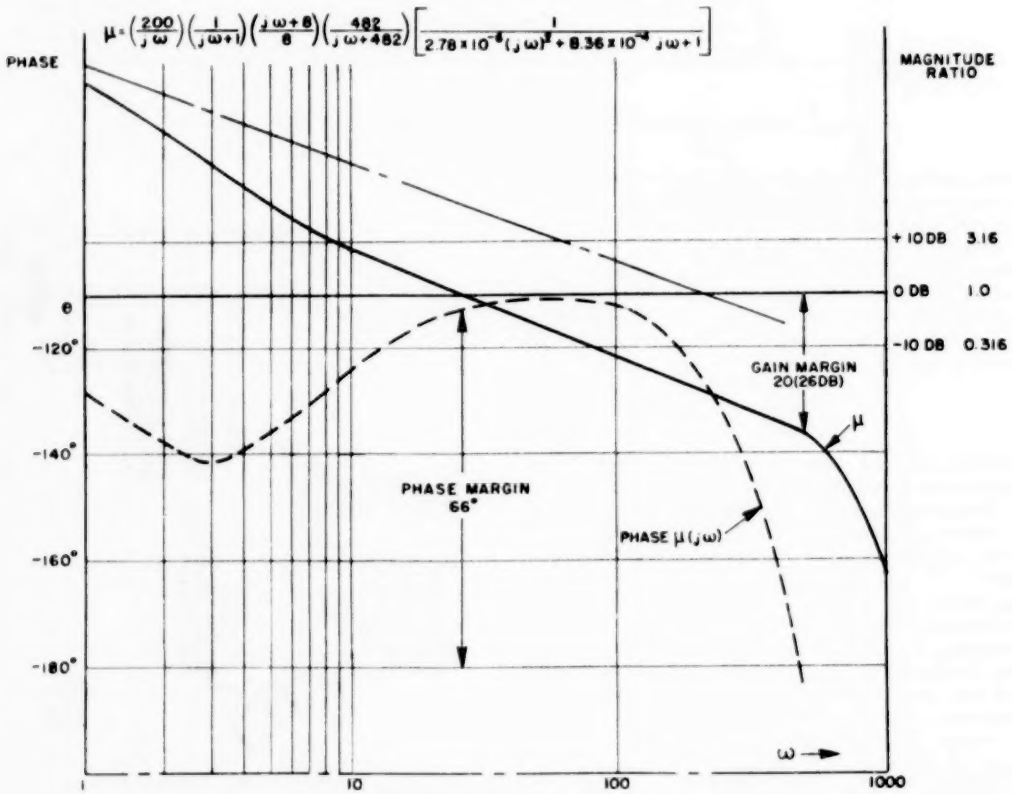


Fig. 14 Frequency-Response Curves for the Final Lathe Design

applied and the resulting velocity it is necessary to integrate the velocity to obtain the displacement of the system; i.e., divide by $j\omega$.

The mechanical impedance of the hydraulic system must now be calculated. This is shown in the Appendix to be

$$\underset{\text{hydraulic}}{Z(j\omega)} = \frac{[MRC(j\omega)^2 + Mj\omega + R]}{j\omega CR + 1} \quad \ldots\ldots [25]$$

where R is A^2/L and $C = V/(BA^2)$.

Substituting Equation [25] in [24] and assuming that only the very low frequencies are of interest so that μ may be considered as $\omega_0/j\omega$

$$\Delta_T = \frac{F_s}{\dfrac{j\omega[MCR(j\omega)^2 + Mj\omega + R]}{j\omega RC + 1} + \dfrac{\omega_0[MRC(j\omega)^2 + Mj\omega + R]}{j\omega RC + 1}} \quad \ldots\ldots [26]$$

For the steady state $\omega = 0$ and the formula for Δ_T becomes

$$\Delta_T = \frac{F_s}{\omega_0 R} \quad \ldots\ldots [27]$$

Taking the calculated value of R as 2×10^2 lb sec/in. from the Appendix and assuming the specified accuracy of 0.0005 in. as Δ_T, the maximum force required to break the static friction F_s which can be tolerated, is 20 lb. This provides a workable margin.

The theoretical design of the servo is now complete. The final step in any design is to build the device and test it. The same techniques of frequency-response analysis can be carried over to the laboratory. The technique consists of applying sinusoidal test signals to the input with the feedback loop open and measuring the amplitude and phase of the output parameter with respect to the input. Such an input in this case can be obtained readily by means of appropriate gearing and a "Scotch yoke" mechanism. The equipment for measuring the gain and phase is more involved and beyond the scope of this paper.

Since there is a large amount of gain in this particular loop, it would be well to measure individual units separately. This laboratory work can well be done at the same time that the analysis is being made. In fact, it is often advantageous to hold up the complete design until a particular element has been evaluated. Almost invariably laboratory measurements will point out factors which have been overlooked in the analysis. This is quite likely to occur in pneumatic systems where the agreement between theoretical and empirical performance is not as exact as, for instance, in electrical systems. The final test is to put the servo together and give it what the electrical engineering profession calls the "smoke test." That is, the power is turned on and the servo operated as a system for the first time.

The frequency-response approach to the design of a servo is an extremely powerful method which enables the designer to work to a stated set of requirements, and to evaluate as the design progresses the adequacy of his components and assemblies. When coupled with empirical data it is also a powerful tool for pointing out errors and deficiencies in the theoretical analysis. Frequency-response methods have had very wide use in the electronic industry for almost two decades. The success of this method has been so great that inevitably the method will extend rapidly to other fields.

Acknowledgment

The author is indebted to Dr. Rufus Oldenburger of the Woodward Governor Company, for his many helpful suggestions and his thorough revision of this paper.

Bibliography

1 "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand Company, Inc., New York, N. Y., 1945.
2 "Transformation Calculus and Electrical Transients," by S. Goldman, Prentice Hall, Inc., New York, N. Y., 1949.
3 "Transients in Linear Systems," by M. F. Gardner and J. L. Barnes, John Wiley & Sons, Inc., New York, N. Y., 1942.
4 "Linear Servo Theory," by R. E. Graham, *Bell System Technical Journal*, vol. 25, 1946, pp. 616–651.
5 "Automatic Feedback Control," by W. Ahrendt and J. F. Taplin, McGraw-Hill Book Company, Inc., New York, N. Y., 1951, p. 273 ff.
6 "Principles of Servomechanisms," by G. Brown and D. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948, p. 136 ff.
7 "Dynamic Behavior of Pneumatic Devices," by L. A. Gould and P. E. Smith, Jr., presented at the Instrument Society of America 7th National Instrument Conference at Cleveland, Ohio, September 8–12, 1952.
8 "Mathematical Engineering Analysis," by R. Oldenburger, The Macmillan Company, New York, N. Y., 1950.
9 "Frequency-Response Data Presentation, Standards and Design Criteria," by R. Oldenburger, published in this issue, pp. 1155–1176.

Appendix

Note on Fig. 1(*k*)

The equation in Fig. 1(*k*) is obtained by assuming that the mass V of air in the region is related to the pressure P_0 in the region by

$$V = CP_0$$

Since

$$Q = \frac{dV}{dt}$$

we have

$$Q = C\frac{dP_0}{dt}$$

Whence

$$P_0 = \frac{1}{C}\int Q dt$$

From Fig. 1(*c*) we have

$$P(t) - P_0 = RQ$$

whence

$$P(t) = RQ + \frac{1}{C}\int Q dt$$

Hydraulic Piston

The response of the hydraulic piston may be derived by noting that the total flow of oil out of the valve must be the sum of the flow due to the movement of the piston, plus the flow due to leakage, plus the flow due to the compressibility of the oil (6). It is assumed that the valve which actuates the piston produces a flow which is proportional to its displacement (and does not introduce a pressure drop across the valve orifices.) Such an "ideal" valve is, of course, impossible to realize, but since the design of components is of secondary interest its use can be justified to simplify the derivation which follows.[6]

[6] Note of R. Oldenburger. The same derivation for rotational motions occurs in Brown and Campbell's book, reference (6), p. 138, and is difficult to justify; the assumptions have not been satisfied in my experience with hydraulic servos.

It is further assumed that the connecting lines also do not introduce a pressure drop. Under these assumptions, the "downstream" pressure in the cylinder will equal the sump pressure which is arbitrarily taken to be zero. Now

$$Q_T = Q_P + Q_L + Q_C$$

where Q_T = total valve flow; Q_P = flow due to movement of piston; Q_L = leakage flow; Q_C = compressibility flow.

In addition

$$Q_P = A \frac{dx_0}{dt}$$

where A = area of piston; x_0 = output displacement of piston.

$$Q_L = LP$$

where L = leakage coefficient; P = pressure drop across piston.

$$\Delta P = \frac{B}{V} \Delta V$$

where B = bulk modulus of oil (0.27×10^6 psi).

$$Q_C = \frac{dV}{dt} = \frac{V}{B} \frac{dP}{dt}$$

where V = total volume of oil under compression. The Δ in the foregoing are increments.

In the relation for compressibility flow, it is assumed that the total volume does not change; i.e., more oil is pumped into the same volume. It is further assumed that the friction in the system is very low and that the loading on the piston is predominantly inertial so that the force opposing the motion of the piston is F where

$$F = M \frac{d^2x_0}{dt^2}$$

The total flow from the valve will be proportional to the valve displacement from zero under the foregoing assumptions so that

$$Q_T = kx_v = A \frac{dx_0}{dt} + LP + \frac{V}{B} \frac{dP}{dt}$$

also

$$P = \frac{F}{A} = \frac{M}{A} \frac{d^2x_0}{dt^2}$$

$$\frac{dP}{dt} = \frac{M}{A} \frac{d^3x_0}{dt^3}$$

so that

$$kx_v = A \frac{dx_0}{dt} + \frac{LM}{A} \frac{d^2x_0}{dt^2} + \frac{VM}{AB} \frac{d^3x_0}{dt^3}$$

In transform notation

$$kx_v(s) = \left(As + \frac{LM}{A} s^2 + \frac{VM}{BA} s^3 \right) x_0(s)$$

and finally

$$\frac{x_0(s)}{x_v(s)} = \frac{k/A}{s\left(\frac{VM}{BA^2} s^2 + \frac{LM}{A^2} s + 1 \right)}$$

We have

$$\frac{x_0(s)}{x_v(s)} = \frac{k/0.136}{s(2.78 \times 10^{-6} s^2 + 8.36 \times 10^{-4} s + 1)}$$

for the following:

Weight of piston and slide assembly = 64 lb
Mass of piston and slide assembly = 2 slugs = M
Effective area of piston (5 in. diam) = 0.136 sq ft = A
Leakage = 7.72×10^{-6} ft³/sec/lb per sq ft = L
Volume = 0.0064 ft³ = V
The valve constant k is determined in the paper.

To evaluate the dynamic effects of external forces the mechanical impedance of the piston must be known. Once the impedance of this member, isolated from the rest of the servo loop, is known, the net effect with feedback is readily calculable. We are assuming that the principle of superposition holds, that is, that the movement of the piston is the resultant (sum) of the movement when there is no external force and the movement when the loop is open, the valve is closed, and the external force is acting on the piston. The impedance may be derived by calculating the resulting velocity when an arbitrary external force is applied to the piston and the valve maintained closed.

As the first step, assume that the mass of the piston and slide is all concentrated at one point external to the piston and that the remainder of the system is massless as shown in Fig. 15. The

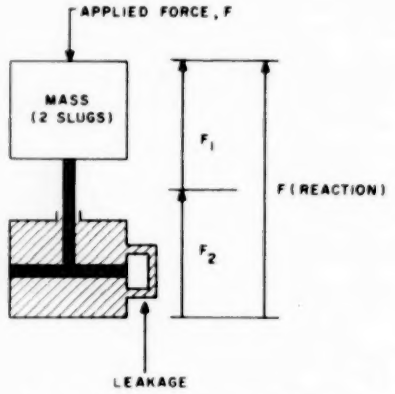


FIG. 15 PISTON AND CYLINDER ASSEMBLY

valve is assumed to be closed. The total applied force F must equal the reaction force from the acceleration of the mass F_1 plus the reaction force F_2 of the piston. Therefore

$$F = F_1 + F_2$$

where

$$F_1 = M \frac{d^2x_0}{dt^2} = M \frac{dx_0}{dt}$$

Clearly F_1 is represented by the transform $F_1(s)$, given by $F_1(s) = Ms\dot{x}_0(s)$. The force F_2 may be evaluated as follows by equating oil flows: The total flow of the oil will be due to the movement of the piston caused by the external force and will be equal to the sum of the leakage and compressibility flows. Now total flow Q_P = leakage flow Q_L + compressibility flow Q_C. Thus we have the following relations when $F(s)$ and $F_2(s)$ are the complex representations of F and F_2, respectively

$$Q_P = A\frac{dx_0}{dt} = A\dot{x}_0$$

$$Q_L = L\frac{F_2}{A} : Q_C = \frac{V}{BA}\frac{dF_2}{dt}$$

$$A\dot{x}_0 = \frac{LF_2}{A} + \frac{V}{BA}\frac{dF_2}{dt}$$

$$A\dot{x}_0(s) = \left(\frac{L}{A} + \frac{V}{BA}s\right)F_2(s)$$

$$F_2(s) = \frac{\dot{x}_0(s)}{\left(\frac{L}{A^2} + \frac{V}{BA^2}s\right)}$$

$$F(s) = F_1(s) + F_2(s) = Ms\dot{x}_0(s) + \frac{\dot{x}_0(s)}{\left(\frac{L}{A^2} + \frac{V}{BA^2}s\right)}$$

Introducing R and C where

$$R = \frac{A^2}{L}, \quad C = \frac{V}{BA^2}$$

we have $$F(s) = Ms\dot{x}_0(s) + \frac{R\dot{x}_0(s)}{(sCR + 1)}$$

whence $$F(s) = \frac{(MRCs^2 + Ms + R)}{(sCR + 1)}\dot{x}_0(s)$$

By definition the mechanical impedance $Z(s)$ is such that

$$F(s) = \dot{x}_0(s)Z(s)$$

whence $$Z(s) = \frac{(MRCs^2 + Ms + R)}{(sCR + 1)}$$

Note of Rufus Oldenburger[2] on Pneumatic Components

We shall establish the validity of Equation [18]. Let Q_J, Q_N, and Q_L denote the rates of flow through the orifice, through the flapper nozzle, and to the output volume in Fig. 7. Let D_s be the diameter of the orifice restriction, P_s the supply pressure, and P_0 the downstream pressure as shown in the figure. For a proportionality constant K_J the flow through the orifice (8) is given by

$$Q_J = K_J D_s^2\sqrt{P_s - P_0}$$

The atmospheric pressure is taken to be zero, whence P_0 and P_s are gage pressures. The flow through the nozzle may be taken to be

$$Q_N = K_N D_N Y\sqrt{P_0}$$

where D_N is the diameter of the nozzle, Y is the distance from the flapper to the nozzle (properly measured), and K_N is a proportionality constant. Let Y_0 be the initial value of Y. The input displacement x_f of the flapper is taken to be $Y_0 - Y$. The reader will note that an increase in x_f implies an increase in the pressure P_0. The quantity Q_L is such that

$$Q_L = Q_J - Q_N$$

In view of the foregoing formulas for Q_J and Q_N we have

$$Q_L = f(P_0, x_f)$$

for a function $f(P_0, x_f)$ of P_0 and x_f.

Taking differentials we have

$$dQ_L = \frac{\partial Q_L}{\partial P_0}dP_0 + \frac{\partial Q_L}{\partial x_f}dx_f$$

To simplify notations we replace dQ_L, dP_0, and dx_f by Q_L, P_0, and x_f, respectively, so that then

$$Q_L = \frac{\partial Q_L}{\partial P_0}P_0 + \frac{\partial Q_L}{\partial x_f}x_f$$

We assume that $$Q_L = C_p\frac{dP_0}{dt}$$

for a capacity constant C_p. We thus have

$$C_p\frac{dP_0}{dt} = \frac{\partial Q_L}{\partial P_0}P_0 + \frac{\partial Q_L}{\partial x_f}x_f$$

We introduce symbols R_s and μ_p so that

$$\frac{\partial Q_L}{\partial P_0} = -\frac{1}{R_s}, \quad \frac{\partial Q_L}{\partial x_f} = \frac{\mu_p}{R_s}$$

We then have

$$C_p\frac{dP_0}{dt} = -\frac{P_0}{R_s} + \frac{\mu_p x_f}{R_s}$$

whence Equation [18] is valid. The quantities R_s and μ_p are by definition resistance and gain constants. It follows that

$$R_s = -\frac{1}{\dfrac{\partial Q_L}{\partial P_0}}, \quad \mu_p = -\frac{\dfrac{\partial Q_L}{\partial x_f}}{\dfrac{\partial Q_L}{\partial P_0}}$$

Derivation of Pneumatic Equations [16] and [17]

For a commercially available amplifier the empirically determined constants are

$$K_J = 3000$$
$$K_N = 9000$$

We introduce β where

$$\beta = D_s^2/D_N$$

From the definition of R_s and μ_p in the preceding section there follows

$$R_s = \frac{2}{3000\,D_s^2(P_s - P_0)^{-1/2} + 9000\,D_N Y P_0^{-1/2}}$$

$$\mu_p = \frac{6P_0^{1/2}}{\beta(P_s - P_0)^{-1/2} + 3YP_0^{-1/2}}$$

We assume that $\sqrt{P_s - P_0}$ is of the same order of magnitude as $\sqrt{P_0}$. Also, we suppose that

$$D_s = D_N$$

and that $$Y < \frac{D_N}{10}$$

It follows that μ_p may be approximated by

$$\mu_p = \frac{6}{D_N}\sqrt{P_0(P_s - P_0)}$$

[2] Director of Research, Woodward Governor Company, Rockford, Ill.

We use the average pressure P_0 for P_0 and assume that $\overline{P_0} = 1/2\ P_s$. Then

$$\mu_p \doteq \frac{6\overline{P_0}}{D_N}$$

In a similar manner with the same assumptions, R_s may be shown to be

$$R_s = \frac{(P_0)^{1/2}}{1500\ D_s^2}$$

whence the Relations [16] are established.

To establish the validity of Equation [17] we note that for Fig. 7

$$Q_L = C\frac{dP_0}{dt} + A_\beta \frac{dx_v}{dt}$$

where C is the capacity of the bellows (output volume in Fig. 7) when the volume of the bellows is fixed, A_β is the area of the moving face of the bellows, x_v is the co-ordinate of this face (identical with the hydraulic-valve co-ordinate), and Q_L is the rate of flow to the bellows. The reader may imagine that the bottom edge of the output volume in Fig. 7 is the moving face of the bellows. The spring force F_3 for the spring attached to the bellows satisfies the formula

$$F_3 = K_s x_v$$

Since

$$F_3 = A_\beta P_0$$

it follows that

$$x_v = \frac{A_\beta P_0}{K_s}$$

and

$$Q_L = \left(C + \frac{A_\beta{}^2}{K_s}\right)\frac{dP_0}{dt}$$

The effective capacity of the bellows is thus given by

$$C_p = \left(C + \frac{A_\beta{}^2}{K_s}\right)$$

Corresponding to the average absolute pressure $P_0 + 15$ psi (here 15 is the atmospheric pressure in psi) we suppose that there is a mass V_1 of air in the bellows so that

$$V_1 = C(P_0 + 15)$$

Relation [17] now follows.

Note of Rufus Oldenburger on Use of a Driving Pressure P_1 in Fig. 11

The use of the driving pressure P_1, where

$$P_1 = \mu_p x_f$$

in Equation [23] will now be justified. We consider Fig. 7 where the output volume of this figure is the capacity of Fig. 11, i.e., the capacity introduced in Fig. 13. The pressure P_0 of Fig. 7 will now be replaced by the symbol P_2 to agree with Fig. 11 and the pressure in the capacity of Fig. 7 will be denoted by P_0. We must imagine that there is an orifice in Fig. 7 just ahead of the output volume so that the pressure P_0 in the capacity is different (at least can be different) from the pressure P_2 on the other side of the orifice (see Fig. 13). Let Q_L be the rate of flow of air to the capacity, Q_J the rate of flow through the orifice drawn in Fig. 7, and Q_N the rate of flow through the nozzle, whence

$$Q_L = Q_J - Q_N$$

as in the previous discussion concerning Fig. 7. Now

$$Q_J = f(P_2), \quad Q_N = g(P_2, x_f)$$

for functions $f(P_2)$ of P_2 and $g(P_2, x_f)$ of P_2 and x_f. It follows that

$$Q_L = h(P_2, x_f)$$

for a function $h(P_2, x_f)$ of P_2 and x_f. Taking differentials

$$dQ_L = \frac{\partial h}{\partial P_2}\,dP_2 + \frac{\partial h}{\partial x_f}\,dx_f$$

To simplify notations, as before we shall replace the differentials by the symbols for the variables themselves and write

$$Q_L = \frac{-P_2}{R_1} + \frac{\mu_p x_f}{R_1}$$

where

$$-\frac{1}{R_1} = \frac{\partial h}{\partial P_2}, \quad \frac{\mu_p}{R_1} = \frac{\partial h}{\partial x_f}$$

Let R_2 be the resistance of the orifice just ahead of the capacity. From the definition of R_2 we have

$$R_2 Q_L = P_2 - P_0$$

Since

$$Q_L = CP_0'$$

this yields

$$R_2 CP_0' + P_0 = P_2$$

whence in complex form

$$P_0(s) = \frac{P_2(s)}{R_2 Cs + 1}$$

Thus in complex form

$$Q_L(s) = \frac{CsP_2(s)}{R_2 Cs + 1}$$

and

$$\frac{CsP_2(s)}{R_2 Cs + 1} = \frac{-P_2(s)}{R_1} + \frac{\mu_p x_f(s)}{R_1}$$

Solving, we obtain

$$P_2(s) = \frac{(R_2 Cs + 1)\mu_p x_f(s)}{[C(R_1 + R_2)s + 1]}$$

With $\mu_p x_f$ replaced by P_1 this yields the bottom relation in Fig. 11; whence the use of the argument in Fig. 11 is correct.

Discussion

R. Oldenburger.[6] The writer believes that each major theoretical approach to the design of feedback loops has its own area of application where it excels over others. This paper is particularly important in that the solution to the problem is readily suggested by the frequency-response curves, whereas the chance that one would think of this solution in working with

[6] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

algebraic or other analytical methods appears to the writer to be very small. According to Elliot Woodhull (Perkin-Elmer Corp.) pneumatic control engineers are familiar with the author's solution, having discovered it by accident.

The frequency-response and algebraic methods will be compared for the author's problem. We shall start with the frequency-response point of view. With the pilot relay in the picture the author has the frequency-response curves of Fig. 10, where the gain and phase margins are woefully inadequate. If we keep the phase curve fixed, but shift the magnitude-ratio curve down, both gain and phase margins are increased, thus improving the stability of the system. Also, if the magnitude curve is held fixed, but the phase curve is shifted up, both gain and phase margins are improved. The problem is thus to introduce a physical component that in the region of interest moves the phase curve up and the magnitude-ratio curve down, improving the margins accordingly. By introducing a capacity as shown in Fig. 13, the author replaces the factor

$$\frac{1}{\frac{j\omega}{145}+1}$$

of the open-loop transfer function by the factor

$$\frac{\frac{j\omega}{8}+1}{j\omega+1}$$

For $1000 \geqq \omega \geqq 30$ this substitution decreases the open loop-magnitude ratio and at the same time increases the phase. For $\omega \gg 1000$ [ω dominating 1000] the phase is increased while the magnitude ratio also is increased. However, both gain and phase margins occur at values of ω definitely less than 1000 so that we need not be concerned with the changes for $\omega \gg 1000$. For $\omega < 30$ both phase and magnitude ratio are decreased by the foregoing substitution. The decrease in magnitude ratio is favorable while the decrease in phase is not. The latter has a small effect on the gain and phase margins. The net effect of introducing the capacity of Fig. 13 is to increase the gain and phase margins considerably.

We shall now turn to the algebraic point of view. We consider a system with the open-loop transfer function μ given by

$$\mu = \frac{K}{\alpha s^3 + \beta s^2 + s} \quad [28]$$

where K, α, and β are positive constants. The characteristic equation is now

$$s^3 + \frac{\beta}{\alpha} s^2 + \frac{s}{\alpha} + \frac{K}{\alpha} = 0 \quad [29]$$

obtained by setting $(\mu + 1)$ equal to zero. We suppose that this equation satisfies the Routh stability condition, namely, that

$$\left(\frac{\beta}{\alpha}\right)\left(\frac{1}{\alpha}\right) \gg \frac{K}{\alpha} \quad [30]$$

Suppose now that Equation [29] has a numerically large negative root. This root will then (see the writer's discussion of the Evans paper published in this issue, pp. 1335–1344) be approximately equal to $-\beta/\alpha$. By the writer's right-to-left division (see the same discussion) we have

1	β/α	$1/\alpha$	K/α	β/α
1	ϵ	K/β		

Here (K/β) is the result of dividing (K/α) by (β/α). Note that by the inequality [30] the quantity $(1/\alpha)$ dominates (K/β) so that (K/β) is the difference of two relatively large quantities. Since one root is approximately $-\beta/\alpha$ the quantity ϵ is relatively small. Multiplying ϵ by the divisor (β/α) we obtain the quantity $(\epsilon\beta/\alpha)$, which must be approximately equal to $(1/\alpha)$. We assume that $\epsilon^2 \ll 4K/\beta$. The roots of Equation [29] are now approximately

$$\frac{-\beta}{\alpha}, \quad \frac{-\epsilon}{2} \pm i\sqrt{\frac{K}{\beta}} \quad [31]$$

where the complex roots correspond to highly oscillatory transients. In the analysis the root $-\beta/\alpha$ can be neglected since the corresponding transient term in the solution dies out rapidly.

Since highly oscillatory transients are undesirable, in so far as stability is concerned, we modify the transfer function μ with the factor

$$\frac{as+1}{bs+1}$$

to obtain μ_1, where

$$\mu_1 = \frac{K(as+1)}{s(bs+1)(\alpha s^2 + \beta s + 1)}$$

We shall choose b so that the lag corresponding to b is large compared to that corresponding to the quadratic factor $(\alpha s^2 + \beta s + 1)$. This quadratic factor can then be dropped from μ_1 to give μ_2, where

$$\mu_2 = \frac{K(as+1)}{s(bs+1)}$$

Thus, if

$$\alpha s^2 + \beta s + 1 = 0$$

has complex roots we assume that β/α dominates $1/b$.

The characteristic equation

$$\mu_2 + 1 = 0$$

now yields

$$bs^2 + (Ka+1)s + K = 0 \quad [32]$$

We shall choose a so that

$$Ka \gg 1 \quad [33]$$

whence the roots of Equation [32] are

$$-\frac{Ka}{2b} \pm \frac{1}{2b}\sqrt{K^2a^2 - 4bK}$$

If

$$b = \frac{Ka^2}{4} \quad [34]$$

the two roots are identical and are equal to

$$\frac{-2}{a}, \quad \frac{-2}{a} \quad [35]$$

The roots in the Expression [31] have been replaced by those of Expression [35]. A system with the roots given in Expression [35] is quite stable. The constant a should be chosen as small as possible so that the roots given in Expression [35] are numerically large and the corresponding physical system is a rapid one while at the same time the inequality in Formula [33] is satisfied.

We shall show that the foregoing algebraic theory underlies the author's work. The factors

$$\frac{145}{j\omega + 145} \quad \frac{482}{j\omega + 482}$$

in the transfer function μ of Fig. 10 correspond to relatively small time lags, in fact to time constants of $^1/_{145}$ and $^1/_{482}$ sec. We therefore drop these factors from μ. The quadratic factor

$$2.78 \times 10^{-6} s^2 + 8.36 \times 10^{-4} s + 1$$

with $s = j\omega$ occurs in the denominator of μ.

This factor yields the roots

$$-150 \pm 580i$$

when set equal to zero. From these roots it can be seen that the time constant $^1/_{145}$ is by no means out of the picture, but can be neglected in a rough first approximation. The function μ now reduces to

$$\mu = \frac{200}{s(3 \times 10^{-6} s^2 + 8 \times 10^{-4} s + 1)}$$

where we have rounded off numbers. Thus

$$K = 200, \alpha = 3 \times 10^{-6}, \beta = 8 \times 10^{-4}$$

The equation $\mu + 1 = 0$ yields

$$s^3 + 300s^2 + 3 \times 10^5 s + 7 \times 10^7 = 0$$

This cubic corresponds to a physical system on the verge of instability (see the Routh Condition [30]). By the writer's right-to-left division the quantity -240 is seen to be a root as shown.

1	300	3×10^5	7×10^7	$\vert 240$
		10^5		
60		29×10^5		

The roots of the reduced equation

$$x^2 + 60x + 29 \times 10^4 = 0$$

are

$$-30 \pm 540i$$

These correspond to highly oscillatory transients.

We multiply μ by the factor

$$\frac{\frac{s}{8} + 1}{s + 1}$$

Then $a = {}^1/_8$, $b = 1$. The factor $s + 1$ corresponds to a very large lag compared to the quadratic factor $(3 \times 10^{-6} s^2 + 8 \times 10^{-4} s + 1)$ whence this quadratic factor can be neglected. The transfer function becomes

$$\mu_2 = \frac{200\left(\frac{s}{8} + 1\right)}{s(s + 1)}$$

The characteristic equation is now

$$s^2 + 26s + 200 = 0$$

with the roots $-13 \pm 6i$. These roots correspond to a very stable system with fast response. Had we taken $a = {}^1/_8$, but b so that Equation [34] is satisfied we would have had $b = {}^{25}/_{32}$, and obtained the roots

$$-16, -16$$

The value $b = 1$ actually used by the author is only slightly different from that just given.

The reader can thus see that the algebraic reasons for the insertion of the capacity in Fig. 13 are rather complicated, and it appears to the writer that only an expert algebraist would be likely to find the solution from this approach.

A more accurate picture can be secured by obtaining the actual characteristic roots. The roots for the function μ of Formula [20] associated with Fig. 4 are

$$-14 \pm 74i \quad [36]$$

The corresponding transients are too oscillatory. Insertion of the pilot relay to obtain Fig. 10 yields the transfer function μ of Formula [22] from which the characteristic equation $\mu + 1 = 0$ with s for $j\omega$ is

$$s^5 + 10^5 s^4 + 6 \times 10^8 s^3 + 2 \times 10^8 s^2 + 2 \times 10^{10} s + 4 \times 10^{12} = 0 \quad [37]$$

To obtain Equation [37] we have rounded off some numbers. To solve this equation we reduce the roots by a factor of 100. This is accomplished by replacing s by $100x$, whence

$$x^5 + 10x^4 + 60x^3 + 200x^2 + 200x + 400 = 0 \quad [38]$$

From the first two terms we have the trial root -10. Taking -5 instead we have the right-to-left division

1	10	60	200	200	400	$\vert 5$
		24	180	120		
	5	36	24	80		

The reduced equation is

$$x^4 + 5x^3 + 36x^2 + 24x + 80 = 0 \quad [39]$$

The first three terms yield the trial factor $x^2 + 5x + 36$ (see the writer's discussion of the Evans paper, pp. 1335–1344) whence the writer's right-to-left division yields

1	5	36	24	80	1, 5, 36
		2	10		1, 0.4, 2
		34	14		
	0.4	2			
	4.6	32			
1	5	36			
	−0.4	−4			

Since the remainders are small the expressions $(x^2 + 5x + 36)$ and $(x^2 + 0.4x + 2)$ are factors of the left side of Equation [39], and hence Equation [38]. Setting these quadratic factors equal to zero we obtain the complex roots of Equation [39]. The roots of Equation [38] are now

$$-5, -2.5 \pm 5.5i, -0.2 \pm 1.4i$$

Multiplying by 100 we have the roots

$$-500, -250 \pm 550i, -20 \pm 140i \quad [40]$$

for Equation [37]. The last two roots in the Expression [40] are really no better than the roots of Formula [36]. We are still on the verge of instability.

For the final configuration of Fig. 13 the characteristic equation, after rounding off numbers, is

$$s^5 + 800s^4 + 5 \times 10^5 s^3 + 2 \times 10^8 s^2 + 5 \times 10^9 s + 4 \times 10^{10} = 0 \quad [41]$$

Replacing s by $100x$ we have

$$x^5 + 8x^4 + 50x^3 + 200x^2 + 50x + 4 = 0 \quad [42]$$

The first two terms yield the trial root —8. Actually —5.3 is a root as shown by the division

1	8	50	200	50	4	5.3
		14	190	49		
1	2.7	36	9.3	0.75		1, 2.4, 35
		0.02	0.05			1, 0.27, 0.021
		36	9.3			
	0.27	0.65				
	2.4	35				
1	2.4	35				

In this division we also have removed the quadratic factor $(x^2 + 2.4x + 35)$ suggested by the first three terms $x^4 + 2.7x^3 + 36x^2$ of the reduced equation obtained as shown after removal of the root —5.3. The other quadratic factor is $(x^2 + 0.27x + 0.021)$. The roots of Equation [42] are thus

$$-5.3, \quad -1.2 \pm 6i, \quad -0.14 \pm 0.17i$$

Multiplying by 100 we obtain the roots

$$-530, \quad -120 \pm 600i, \quad -14 \pm 17i \quad [43]$$

The dominant roots $-14 \pm 17i$ for Fig. 13 are completely satisfactory whereas the corresponding roots $-20 \pm 140i$ of Expression [40], when the capacity is omitted, are quite unsatisfactory.

From the first four coefficients of Equation [42] we form the first Routh difference

$$\left(50 - \frac{200}{8}\right) \quad [44]$$

The first term in this difference is only twice the second term. The writer normally tries to design for at least a 3:1 or 4:1 ratio of these quantities. Many other control people do the same. One might conclude that although the author's solution is not critical as to the value of the gain constant, it may be critical as to the other constants. Actual analysis, which we omit due to limitations of space, shows that wide variations can be allowed in the other constants, and that they can be chosen so that the ratio of the first term to the second can even be less than 1.2, in which case the terms in Formula [44] are almost identical, while the system is still quite stable. This is a remarkable property of the author's device and emphasizes the importance of obtaining the actual characteristic roots in contrast to relying on stability criteria.

We shall now turn to the Appendix. In his Appendix the author assumes that there are no pressure drops across the orifices of the hydraulic valve supplying the hydraulic cylinder. In hydraulic servo work we generally assume that the pressure drop across the piston is negligible, and that the pressure drop across the intake ports equals this drop for the discharge ports. We generally design our servos so that these properties are satisfied for practical purposes. Let P_0 and P_a denote supply and sump pressures, respectively. The law relating x_0 and x_v is now

$$A\frac{dx_0}{dt} = ax_v\sqrt{\frac{P_0 - P_a}{2}} \quad [45]$$

for a constant a. With P_0 and P_a constant the speed of the piston is then proportional to valve opening, at least for rectangular ports.

When there is an appreciable load on the piston, Equation [45] no longer holds. Let F be the force on the piston due to all causes except for the pressure drop across the piston and its inertia. Also, let P_1 and P_2 be the pressures in the cylinder on the upstream and downstream sides of the piston, respectively. Finally, let V_1 and V_2 be the volumes of oil under compression in the cylinder on the upstream and downstream sides of the piston, respectively. Assuming identical flow characteristics for the intake and discharge ports the following equations hold.

$$\left.\begin{aligned} ax_v\sqrt{P_0 - P_1} &= Ax_0' + \frac{V_1}{B}P_1' + L(P_1 - P_2) \\ F + A(P_1 - P_2) &= M\frac{d^2x_0}{dt^2} \\ ax_v\sqrt{P_0 - P_1} &= \frac{V_1P_1'}{B_1} + \frac{V_2P_2'}{B} + ax_v\sqrt{P_2 - P_a} \end{aligned}\right\} \quad [46]$$

Here the primes denote derivatives with respect to time. From these equations P_1 and P_2 can be eliminated (theoretically at least) yielding an equation in

$$x_v, F, \text{ and } x_0$$

To have this reduce to a relation between x_v and x_0 the function F must be given. Equations [46] are correct servo equations.

The friction force F_s can be computed rigorously as follows: It is assumed that the servo piston is not moving, being opposed by a friction force F_s while the static error Δ_T is 0.0005 in. Here Δ_T is $x_{in} - x_0$. See Fig. 3. With the notations of the Appendix we have

$$F_s = AP \quad [47]$$

$$kx_v = LP \quad [48]$$

Equation [47] arises from the assumption that the only forces on the piston mechanism are the friction force F_s and the force due to the pressure drop across the piston. Equation [48] is true because in steady state no oil is in the process of being stored in the cylinder, and the rate of oil flow through the valve is equal to the rate of oil flow past the piston due to leakage. From Equations [47] and [48] one has

$$F_s = \frac{kAx_v}{L} \quad [49]$$

In steady state the valve opening x_v and the error Δ_T are related by

$$x_v = 100\Delta_T \quad [50]$$

for the numerical case under study. With

$$A = 0.136 \text{ ft}^2$$
$$k = 0.023 \text{ ft}^3/\text{sec/in.}$$
$$L = 7.72 \times 10^{-4} \text{ ft}^3/\text{sec/lb per sq ft}$$

Equations [49] and [50] yield

$$F_s = 4 \times 10^4\Delta_T$$

With $\Delta_T = 0.0005$ in. this gives 20 lb for F_s. This agrees with the figure of 20 lb obtained by the author using an entirely different theory.

In his discussion above Fig. 14 the author states, "there is no loss in the network or pilot relay for steady pressures." This statement is not strictly correct since the diaphragm in the pilot relay can be in different steady-state positions depending on the values of the pressures. The pressure drop across the diaphragm is zero only if the diaphragm is in the neutral unstrained position. Suppose that we start with a steady state in which the diaphragm

is in the neutral position. Now move the flapper a given distance toward the closed position. The input pressure P_0 to the pilot relay (see Fig. 9) will then increase forcing the diaphragm toward the nozzle so as to increase the transmitted relay pressure P_r to a new value. Since in its new steady-state position the diaphragm is strained, the transmitted pressure is less than the input pressure. In fact, if the diaphragm is 0.001 in. off from the neutral position the pressure drop across the diaphragm is 0.3 psi.

E. H. Woodhull.[7] The author has done a very fine job of presenting in a systematic manner the analytical considerations involved in the design of a mechanical servo. He has introduced the necessary components first and then added networks as required to meet the design condition. The paper also presents a very interesting combination of pneumatic and hydraulic components in the same device.

Of particular interest is the derivation of the nozzle-baffle-orifice transfer function in a manner analogous to that used to obtain triode characteristics. It should be pointed out that experimental measurements of flow versus nozzle back pressure with constant nozzle-baffle gap can be obtained and the incremental resistance and transconductance can be determined precisely for any operating pressure. It is particularly important to obtain an experimental determination for the case of restrictions which are neither orifices nor capillaries, but something in between.

Also of considerable interest is the derivation of the transfer function of the pneumatic circuit shown in Fig. 11, which is a very useful pneumatic device for obtaining integral equalization. The pneumatic resistances shown in the circuit of Fig. 13 are orifices which establish a square-root relationship between pressure differential and flow and hence are nonlinear resistors and can be treated as linear only for small increments about an equilibrium position. On the other hand, a capillary which follows Poiseulle's law is a constant pneumatic resistance as long as the average pressure is held constant. If orifices are to be used as pneumatic resistances, it is wise to evaluate the circuit stability at the pressure extremes which are encountered in normal operation of the circuit, since the time constants can vary considerably.

Author's Closure

Dr. Oldenburger's algebraic solution of the servo problem is a very fine work of *analysis*. However, only an expert algebraist would be capable of the *synthesis* of a servo from the basic requirements by his methods. The great virtue of the frequency-response method is that it offers a systematic approach to the *synthesis* problem that is readily mastered. In practice, a completely algebraic approach could possibly degenerate into a matter of successive approximations which may or may not converge to an acceptable solution. Although the servo used as an example in the paper is not by any means a complicated one, the characteristic equation is of the fifth order. More complicated problems or problems in which the smaller parasitic effects are included have characteristic equations of correspondingly higher order. This makes any algebraic method extremely laborious. In addition, graphical methods seem to lend themselves more readily to general engineering use.

Any test for stability may test for singularities in the right half of the complex frequency plane or examine the roots of the characteristic equation for positive real parts. The frequency-response methods are based upon the stability criteria due to Nyquist. This criteria is simply a test for singularities in the right half of the complex frequency plane. Routh's criteria is exactly the same type of test. As a matter of fact the derivations of both criteria are based upon the same theorem of Cauchy's. For this reason the analysis of the effects due to variations in the parameters of the system may be accomplished by either the algebraic method due to Routh or the topographical method due to Nyquist. The writer prefers the frequency-response method as he feels there is a much closer relation between the analysis and the actual physical makeup of the servo. It is rather easy to determine the critical parameters. For a simple device such as is under discussion here a complete replot of the frequency response can be made in half an hour. Thus, in a very short time the effect of variations is made evident. The limited scope of the paper did not permit of any detailed discussion of the effects of such variations. Although perhaps not stated explicitly the design objective was to meet the design requirements with ample margin for variation of the parameters of the servo. As Dr. Oldenburger states the servo does have wide tolerance to variations. However, this is not due to the particular method by which the servo is stabilized. Any servo whose open-loop frequency response is the same as Fig. 14 will exhibit the same tolerance to variations.

The derivation of the response curves of the hydraulic valve and piston was highly simplified. A rigorous derivation would form an entire paper by itself. The derivation in reference (6) is for a positive displacement pump with variable stroke controlling the flow of oil rather than a variable orifice valve. For a positive displacement pump the flow is proportional to the length of stroke and the speed of the pump and is independent of the pressure drop across the pump piston. It was felt that emphasis should be placed more upon the manipulation of the frequency-response curves than upon a rigorous description of a component. For this reason a simplified derivation that is based upon the derivation in reference (6) was incorporated in the paper. When the results agreed with published curves of commercially available valves and cylinders the writer felt that the accuracy of the approximation was sufficient for the purpose intended.

The methods of frequency-response analysis can be applied only to systems which can be described by *linear* differential equations with constant coefficients. Dr. Oldenburger's Equations [46] are not linear and, hence, can not be used without change. To utilize the more exact [46] it would be necessary to compute the effect of small variations in exactly the same way as was done for the case of pneumatic elements where the same equations apply. This, in effect, linearizes the equations.

It must be admitted that the derivation of the effect of external forces upon the accuracy of the servo in the paper is considerably more complicated that Dr. Oldenburger's. However, the derivation in the paper is the solution for the general case in which the effect of any arbitrary applied external force can be calculated. Dr. Oldenburger's derivation holds only for a time invariant force. Equation [26] can be used as a transform, however, to calculate the effects of dynamic forces such as vibration.

The pressure drop across the pilot relay may be computed from Equation [21] for the steady state by setting $\omega = 0$. For the constants assumed in the paper $P_r = 0.94\ P_0$, or about a 6 per cent pressure drop. In view of the wide tolerance of the servo to variations in static gain it simply did not seem worth while to bother with the 6 per cent at the time the paper was written.

The writer is in complete agreement with Mr. Woodhull's comments and would also like to emphasize the importance of evaluating empirically the performance of the servo at the extremes of the supply pressures, both pneumatic and hydraulic.

The writer would like to extend his most sincere thanks to Dr. Oldenburger and the ASME for their generous assistance in the preparation of this paper.

[7] Development Engineer, The Perkin-Elmer Corporation, Norwalk, Conn. Mem. ASME.

Determination of Transient Response From Frequency Response

By A. LEONHARD,[1] STUTTGART, GERMANY

When calculating the transient response of a control loop by means of the classical differential-equation method, the solution of the characteristic equation and the determination of the integration constants becomes extremely difficult for complicated configurations. If the operational calculus, i.e., the Laplace transformation, is applied, the constants of integration are produced automatically as long as definite but meaningful initial conditions are assumed, and the calculating work is simplified considerably. Even with this method, however, the roots of the characteristic equation must still be found. This presupposes that the equation can be expressed analytically, and that it can be solved, which is not always the case, as for example, if the equation is transcendental. Graphical methods have been developed for such cases, which allow the determination of the transient output response with sufficient approximation and with a moderate amount of calculating work. These methods, based on the frequency response, will be treated in detail in this paper, and their practical application will be demonstrated by means of a few problems.

Introduction

In the treatment of control problems, the methods of the operational calculus, i.e., the Laplace transformation, have become more and more common during the past decade. The basic reason for this has been the fact that the solution of increasingly complicated problems had to be and was attempted. The classical method, employing only the differential equation, either failed completely for such problems, or required an unreasonably large amount of calculating work. For systems described by a differential equation of higher than third order, the determination of the constants of integration especially becomes very troublesome when using the classical method.

When using the operational calculus and the Laplace transformation, this troublesome determination of the integration constants becomes unnecessary if we restrict ourselves to such transients which result from the disturbance of a system previously at rest. However, the determination of the roots of the characteristic equation of the differential equation describing the transient response must still be carried out.

It therefore becomes necessary that the equation can be expressed analytically and that it can be solved. For systems containing pure dead time, a transcendental equation with an infinite number of roots results, which cannot be solved even with calculating machines. In such cases, we usually limit ourselves to the determination of the predominant root of the system.

Frequently it is attempted to solve difficult control problems in an experimental fashion by constructing models which represent the entire closed-loop system or individual components of it, and by then conducting measurements on these models. The so-called transfer function, of which more will be said in the following sections, can be determined in this way. The representation of an experimentally plotted curve by an analytical equation which would be necessary in order to derive the equation for the system is, as experience has shown, usually very difficult. Generally, the plotted curve can be approximated only very roughly, and thus the value of the experimental effort is lost again for the greatest part.

As will be demonstrated in the following section, for such cases where the determination of the roots is difficult or even impossible, practical graphical methods are available, which permit the determination of the transient response with a moderate amount of effort.

Determination of Transient Response by Means of Operational Calculus

As shown in Fig. 1, a closed-loop control system S is to be

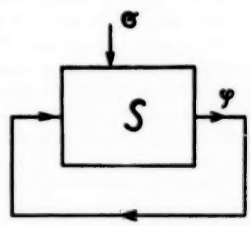


Fig. 1 Closed-Loop Control System

affected by its output, the controlled variable φ, in such a manner that φ assumes a definite desired value, which is to be as independent as possible of system disturbances. Furthermore, a disturbance variable σ shall act upon the system. Both φ and σ are defined as dimensionless deviations from their corresponding steady-state values. If we assume at first that σ varies sinusoidally with time, then all the system variables affected by the disturbance, including the controlled variable temporarily designated as φ_σ, will oscillate sinusoidally with the same frequency as the disturbance, assuming a linear system.

We shall place asterisks on the symbols denoting variables when we use their complex representations. For example, let t be time, $j = \sqrt{-1}$, and ω and σ_0 be given constants. If $\sigma_0 e^{j\omega t}$ represents the disturbance σ, we shall replace $\sigma_0 e^{j\omega t}$ by σ^*.

Consider the open-loop system; i.e., do not feed φ back into the system. The quotient of the output φ_σ^* by the disturbance (input) σ^* is the transfer function relating φ and σ. This will be called the "disturbance transfer function."

Hence we obtain

$$F_\sigma = \frac{\varphi_\sigma^*}{\sigma^*} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [1]$$

for the transfer function F_σ. If, with $\sigma = 0$ and the control loop open, a sinusoidal input φ_f is applied at the same point at which the controlled variable is ordinarily fed back into the system,

---

[1] Professor of Electrical Engineering, Stuttgart Institute of Technology.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 9, 1953. Paper No. 53—A-14.

then the resulting output variable, temporarily denoted as φ_R, also will vary sinusoidally, and the vector ratio of φ_R^* to φ_f^*, called the *open-loop transfer function*, is

$$F_R = \frac{\varphi_R^*}{\varphi_f^*} \quad [2]$$

During ordinary operation, that is, when the loop is closed, σ as well as φ_f act on the system, with φ_f now being equal to φ, and we get

$$\varphi^* = \varphi_\sigma^* + \varphi_R^* = F_\sigma \sigma^* + F_R \varphi^* \quad [3]$$

from which we obtain

$$\frac{\varphi^*}{\sigma^*} = \frac{F_\sigma}{1 - F_R} = F \quad [4]$$

for a function F. Since both F_σ and F_R are functions of $j\omega$, we can replace $j\omega$ by the complex variable s, so that F assumes the form

$$F = \frac{N(s)}{G(s)} = \Phi(s) \quad [5]$$

for a function $\Phi(s)$, and polynomials $N(s)$, $G(s)$, it being assumed that F is a rational function. The vector ratio F is designated as the closed-loop transfer function and it describes the behavior of the system for a sinusoidal disturbance. Even though the response of a system to a sinusoidal disturbance might be of interest in some cases, the response of the system to an arbitrary disturbance is generally of much greater significance. The most troublesome, and therefore the most important type of disturbance is the step-function disturbance. Fortunately, the Laplace transformation offers the possibility of obtaining the response to a step disturbance from the frequency response in a relatively simple manner.

For a step disturbance σ_{ST}, transfer function F as given by Equation [5] divided by s is the Laplace transform of the transient output response $\varphi(t)$. Using the notation of K. W. Wagner (1),[1] this factor disappears. By means of tables, or by applying the rules of the Laplace transformation, the required transient response can be determined from the frequency response.

The use of the Heaviside expansion formula has shown itself to be useful in control problems. With

$$F = \frac{N(s)}{G(s)}$$

(Equation [5]) denoting the closed-loop frequency response, we obtain by means of the expansion

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{N(0)}{G(0)} - \sum \frac{N(s_\nu)}{s_\nu G'(s_\nu)} e^{s_\nu t} \quad [6]$$

where $\nu = 1, 2, \ldots, m$. Here $N(0)$ and $G(0)$ stand for the respective values at $s = 0$, while

$$G'(s_\nu) = \left[\frac{dG(s)}{ds}\right]_{s=s_\nu}$$

and s_ν represent the roots assumed to be distinct, of the equation $G(s) = 0$, which corresponds to the characteristic equation of the differential equation describing the transient response. It is obvious that, when applying the expansion formula as well as when using any other method for determining the time function $\varphi(t)$ from the s-domain, the determination of the roots of the characteristic equation becomes necessary. As already stated, this presupposes that the characteristic equation can be formulated analytically, and furthermore, that it can be solved at all. For all cases where this condition is not fulfilled, and also where the characteristic equation becomes of higher order and calculating machines are not available for its solution, it becomes necessary to accomplish the conversion from frequency to transient response by means of graphical methods, which will be discussed in detail in the following section.

Graphical Determination of Transient Response

It should be stated that the same presuppositions that apply to the use of the Laplace transformation also apply when using the graphical methods. As experience has shown, however, these requirements are always met in control problems. Only the stability must be checked for each individual problem, which corresponds to the determination of whether or not the poles of $\Phi(s)$ or the zero locations of $G(s)$ are all located in the left half of the complex plane.

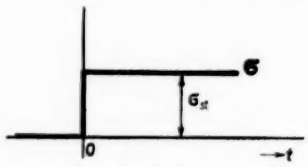


Fig. 2 Step Function

Let us first recall that a so-called step function $\sigma(t)$, as shown in Fig. 2, can be represented as follows

$$\frac{\sigma(t)}{\sigma_{ST}} = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\sin \omega t}{\omega} d\omega \quad [7]$$

According to Equation [7], therefore, the step function of Fig. 2 can be interpreted as a sum of undamped sinusoidal oscillations containing all frequencies between zero and infinity, and having "infinitely" small amplitudes

$$\frac{1}{\pi} \frac{d\omega}{\omega}$$

which, with increasing frequency still continue to decrease.

The transfer function as given by Equation [5], i.e., the behavior of the controlled variable φ for a sinusoidal input σ, is valid for each frequency, and consequently applies to each one of the frequencies of oscillation into which the step disturbance was broken by Equation [7].

If the transfer function $F(j\omega)$, which is a complex number for each value of ω, is written in the form

$$F(j\omega) = A(\omega) e^{j\alpha(\omega)} \quad [8]$$

where A is the amplitude ratio of output to input, and α the phase shift between both quantities at the frequency ω, then it becomes evident that the output response can be written

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \int_0^\infty A(\omega) \frac{\sin[\omega t + \alpha(\omega)]}{\omega} d\omega \quad [9]$$

Each individual oscillation component

$$\frac{d\omega}{\omega} \sin \omega t$$

[1] Numbers in parentheses refer to the Bibliography at the end of the paper.

can be considered as an input oscillation (disturbance σ) and will then appear at the output (controlled variable φ) multiplied by the amplitude ratio $A(\omega)$ and shifted by the phase angle $\alpha(\omega)$, as determined by the frequency response.

As shown in reference (2), the right-hand integral can be approximated immediately by means of a graphical construction. It will be assumed that neither $A(\omega)$ nor $\alpha(\omega)$ change appreciably within a frequency segment $\Delta\omega$ of sufficiently small width, and also, that outside of a specified frequency range $\omega_a < \omega < \omega_b$, the value $A(\omega)$ becomes practically zero. Consequently, a summation can be substituted for the integral, and we obtain

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \sum_{\nu=0}^{\frac{\omega_b - \omega_a}{\Delta\omega}} A(\omega_\nu) \frac{\sin [\omega_\nu t + \alpha(\omega_\nu)]}{\omega_\nu} \Delta\omega \quad [10]$$

where

$$\omega_\nu = \omega_a + \nu\Delta\omega \quad [11]$$

Thus, the approximate output response is not made up any more out of an infinite number of sinusoidal oscillations, but out of a finite and distinct number of such oscillations. The closer the foregoing requirements for $\Delta\omega$ and ω are satisfied, the better will the approximation of the exact response become.

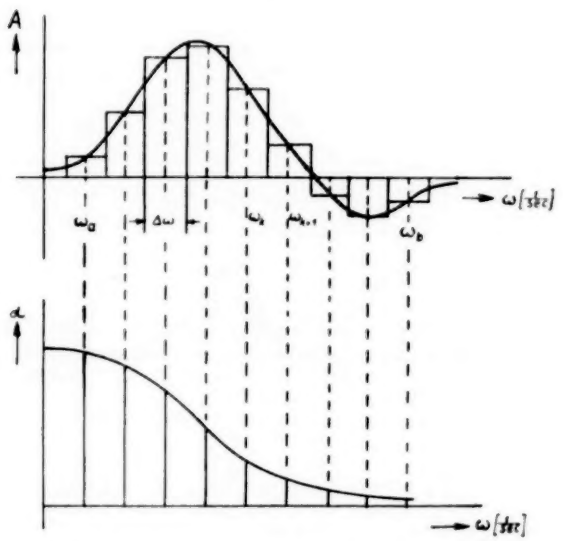


FIG. 3 CLOSED-LOOP FREQUENCY-RESPONSE CURVES $A(\omega)$ and $\alpha(\omega)$ DIVIDED INTO AN APPROPRIATE NUMBER OF SECTORS

Fig. 3, for example, shows the character of $A(\omega)$ and $\alpha(\omega)$ for a specific frequency response $F(j\omega)_0$ divided into an appropriate number of sectors. In order to obtain the value of $\varphi(t_1)/\sigma_{ST}$ for a specific time t_1, the following summation must be formed

$$\sum_{\nu=0}^{\frac{\omega_b - \omega_a}{\Delta\omega}} \frac{A(\omega_\nu)}{\omega_\nu} \sin [\omega_\nu t_1 + \alpha(\omega_\nu)] \Delta\omega$$

The numerical evaluation can be carried out either by means of tabulations, or else graphically as shown in Fig. 4. Circles with the radius

$$\frac{A(\omega_\nu)\Delta\omega}{\omega_\nu}$$

can be drawn for the different frequencies ω_ν, and a time scale corresponding to

$$t = \frac{\beta - \alpha(\omega_\nu)}{\omega_\nu}$$

can be plotted on the circumference of the circles for the different angles β. The values for

$$\frac{A(\omega_\nu)\Delta\omega}{\omega_\nu} \sin [\omega_\nu t + \alpha(\omega_\nu)]$$

can then be taken from these circles for any specific time t_2 (see Fig. 4) and finally all values can be added. Thus the value of the

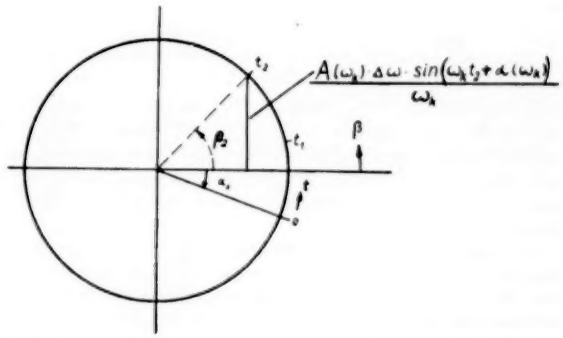


FIG. 4 AUXILIARY CIRCLE FOR DETERMINATION OF $A(\omega_4)\Delta$ SIN $[\omega_4 t + \alpha(\omega_4)]$ FOR ω_4 AND DIFFERENT VALUES OF t

curve for $t = t_2$ has been found. Once the circles have been drawn, additional points of the curve can be found rapidly.

Equation [9] however also can be transformed as follows

$$\begin{aligned}\frac{\varphi(t)}{\sigma_{ST}} &= \frac{1}{2} A(0) + \frac{1}{\pi} \int_0^\infty \frac{A(\omega) \sin [\omega t + \alpha(\omega)]}{\omega} d\omega \\ &= \frac{1}{2} A(0) + \frac{1}{\pi} \int_0^\infty \frac{A(\omega) \cos \alpha(\omega)}{\omega} \sin \omega t d\omega \\ &\quad + \frac{1}{\pi} \int_0^\infty \frac{A(\omega) \sin \alpha(\omega)}{\omega} \cos \omega t d\omega \\ &= \frac{1}{2} A(0) + \frac{1}{\pi} \int_0^\infty \frac{\mathrm{Re}\, F(j\omega)}{\omega} \sin \omega t d\omega \\ &\quad + \frac{1}{\pi} \int_0^\infty \frac{Im\, F(j\omega)}{\omega} \cos \omega t d\omega \quad \ldots\ldots\ldots [12]\end{aligned}$$

Since the step σ_{ST} does not appear prior to $t = 0$, and since the system has previously been at steady state, we can set $\varphi(t) = 0$ for $t < 0$ and therefore obtain [replacing t by $-t$ in Equation [12]]

$$\begin{aligned}0 &= \frac{1}{2} A(0) - \frac{1}{\pi} \int_0^\infty \frac{\mathrm{Re}\, F(j\omega)}{\omega} \sin \omega t d\omega \\ &\quad + \frac{1}{\pi} \int_0^\infty \frac{Im\, F(j\omega)}{\omega} \cos \omega t d\omega \quad [13]\end{aligned}$$

for $t > 0$.

Subsequently for $t > 0$

$$\frac{1}{\pi}\int_0^\infty \frac{Im\ F(j\omega)}{\omega}\cos\omega t d\omega = \frac{1}{\pi}\int_0^\infty \frac{\mathrm{Re}\ F(j\omega)}{\omega}\sin\omega t d\omega - \frac{1}{2}A(0)$$

$$\text{and } \frac{1}{\pi}\int_0^\infty \frac{\mathrm{Re}\ F(j\omega)}{\omega}\sin\omega t d\omega = \frac{1}{\pi}\int_0^\infty \frac{Im\ F(j\omega)}{\omega}\cos\omega t d\omega + \frac{1}{2}A(0)$$

and therefore we obtain

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi}\int_0^\infty \frac{\mathrm{Re}\ F(j\omega)}{\omega}\sin\omega t d\omega \quad [14]$$

or

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi}\int_0^\infty \frac{Im\ F(j\omega)}{\omega}\cos\omega t d\omega \quad [15]$$

Thus, the transient output response can be described either by the real or by the imaginary component of the function $F(j\omega)$, the closed-loop transfer function of the system. This result was already reached by Grünwald in 1941 (3), although he reached the result by a somewhat different method from the one shown in the foregoing.

As compared to the graphical evaluation of the integral of Equation [9], the evaluation of the integrals of Equations [14] and [15] offers the advantage that mechanical calculating devices can be employed to a great extent. At first, it might be mentioned that the method shown for the evaluation of the integral of Equation [9] can, of course, also be applied in this case, with the problem becoming a simpler one because the time scale on all the auxiliary circles, as shown in Fig. 4, now begins at zero for $\beta = 0$.

Equation [10] is now replaced by

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi}\sum_{\nu=0}^{\frac{\omega_b-\omega_a}{\Delta\omega}} \frac{\mathrm{Re}\ F(j\omega_\nu)}{\omega_\nu}\sin\omega_\nu t\Delta\omega \quad [16a]$$

and

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi}\sum_{\nu=0}^{\frac{\omega_b-\omega_a}{\Delta\omega}} \frac{Im\ F(j\omega_\nu)}{\omega_\nu}\cos\omega_\nu t\Delta\omega \quad [16b]$$

If we integrate only from 0 to Ω, where Ω is selected so that Re $F(j\omega)$ or $Im\ F(j\omega)$ has practically reached zero, then the integrals of Equations [14] and [15] correspond to those for the determination of the Fourier coefficients of periodic functions. The various familiar methods for harmonic analysis thus can also be applied to the present problem. A harmonic analyzer, for example, can be used.

The treatment of practical control problems with a harmonic analyzer has, however, shown that the evaluation becomes very troublesome and time-consuming, especially in problems where a number of different oscillations are superimposed on one another, and that, furthermore, the accuracy of the results leaves something to be desired. Since, furthermore, a harmonic analyzer is usually not available to the practical engineer, it will now be demonstrated how, with a tolerable amount of time and effort, the transient output response can be determined from the curves Re $F(j\omega)$ or $Im\ F(j\omega)$. A similar method was developed by J. Hänny (4).

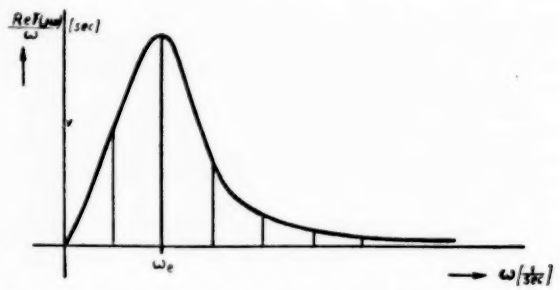


FIG. 5 Re $F(j\omega)/\omega$ FOR A SPECIFIC OUTPUT RESPONSE WITH DISTINCT MAXIMUM

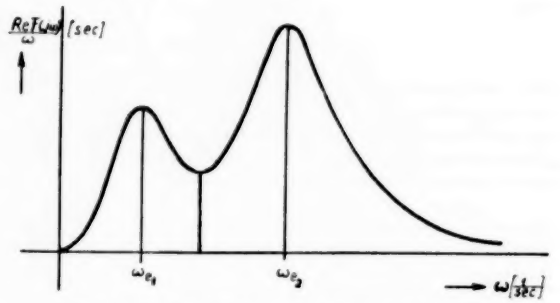


FIG. 6 Re $F(j\omega)/\omega$ WITH TWO DISTINCT MAXIMA

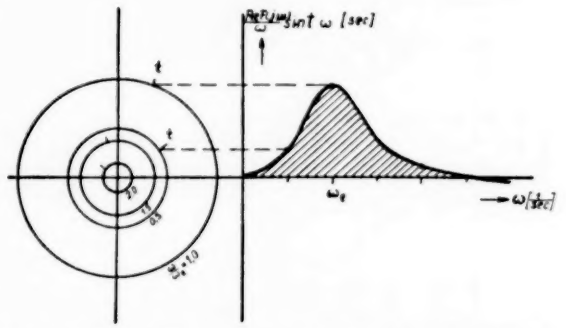


FIG. 7 GRAPHICAL EVALUATION OF ONE POINT ON OUTPUT TRANSIENT RESPONSE BY MEASURING CROSS-HATCHED AREA WITH PLANIMETER

Assume, for example, that the curve Re $F(j\omega)/\omega$ of Fig. 5, is given. This curve has a distinct maximum at the frequency ω_e, which is an indication that an oscillation with this frequency will predominate in the transient response. Since there is a predominant oscillation in most transient responses, the actual curves will usually have a shape similar to that of Fig. 5. If two or more distinct maxima appear, corresponding to two distinct oscillations in the transient response, then the curve is more appropriately subdivided into two or more sections, as shown in Fig. 6. Details on the evaluation of this case will be mentioned later.

We now select a few equally distributed frequencies over the entire frequency range of Fig. 5, the frequencies being either an integral multiple or an integral fraction of the basic frequency. As shown in Fig. 7, we now construct concentric circles with radii

equal to the ordinates Re $F(j\omega)/\omega$ corresponding to the individual frequencies ω/ω_s.

For each value of time t we now plot

$$\frac{\text{Re } F(j\omega)}{\omega} \sin t\omega$$

as a function of ω or ω/ω_s. Using the circles to project the respective ordinates, this procedure can be carried out easily and rapidly. We draw a curve through the points obtained. If we now use a planimeter to measure the area enclosed by the curve obtained above and the ω-axis, then this area for each value of t will be

$$\frac{\pi\varphi(t)}{2\sigma_{ST}} \text{ approx.}$$

If the curve Re $F(j\omega)/\omega$ does not produce one absolute maximum, then ω_s can of course be selected at will. As already mentioned, if the curve, for example, has the shape shown in Fig. 6, it may be appropriate to subdivide it into two or more sections. When the shape of the transient output curve $\varphi_i(t)/\sigma_{ST}$ has been determined for each individual section, then the over-all curve $\varphi(t)/\sigma_{ST}$ results very simply by merely superimposing the different sectional curves $\varphi_i(t)/\sigma_{ST}$.

In some cases, a section of the curve Re $F(j\omega)/\omega$ can be approximated by a straight line. The graphical integration for this section can then be dispensed with, and the corresponding sectional curve $\varphi_i(t)/\sigma_{ST}$ can be found by calculation. If, for example, the section of the curve between the values of ω_1 and ω_2 in Fig. 8 can be replaced by a straight line, then we obtain

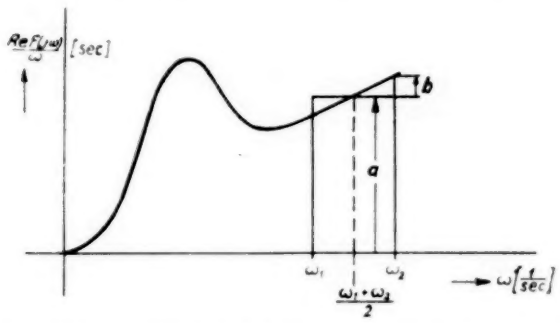


FIG. 8. APPROXIMATION OF A SECTION OF CURVE Re $F(j\omega)/\omega$ BY A STRAIGHT LINE

$$\frac{2}{\pi}\int_{\omega_1}^{\omega_2} \frac{\text{Re } F(j\omega)}{\omega} \sin \omega t d\omega = \frac{2}{\pi}\left[\frac{\cos \omega_1 t}{t}(a-b) - \frac{\cos \omega_2 t}{t}(a+b) + \frac{2b}{(\omega_2-\omega_1)t^2}(\sin \omega_2 t - \sin \omega_1 t)\right] \quad [17a]$$

or, if we are evaluating $Im\ F(j\omega)/\omega$

$$\frac{2}{\pi}\int_{\omega_1}^{\omega_2} \frac{Im\ F(j\omega)}{\omega} \cos \omega t d\omega = \frac{2}{\pi}\left[\frac{\sin \omega_2 t}{t}(a+b) - \frac{\sin \omega_1 t}{t}(a-b) + \frac{2b}{(\omega_2-\omega_1)t^2}(\cos \omega_2 t - \cos \omega_1 t)\right] \quad [17b]$$

It has also been proposed to approximate the curve Re $F(j\omega)/\omega$ by trapezoids, because, with this approximation, the evaluation of the integrals can be conducted relatively easily by means of tables (5).

When employing the graphical method of determining the transient output, we can use either the real or the imaginary component of the transfer function. Generally that curve will be selected which, with increasing values of ω, approaches zero most rapidly, so that the frequency range that must be considered will be the smallest. If the transient output is such that for $t = 0$ the response is not zero but has some initial fixed value, and if values of $\varphi(t)$ for $t \approx 0$ are of interest, then the imaginary component of the transfer function should appropriately be employed. The area between this curve and the ω-axis is then immediately a measure of $\varphi(0)$.

In case a proportional control system is considered, then $A(0) \neq 0$, and also Re $F(0) = A(0) \neq 0$, and consequently

$$\left[\frac{\text{Re } F(j\omega)}{\omega}\right]_{\omega=0} = \infty$$

(One should understand that a limit is taken here.) Thus the curve Re $F(j\omega)/\omega$ cannot be employed for the graphical evaluation. In this case, the curve $Im\ F(j\omega)/\omega$ can be used. If, however, this curve approaches zero slowly for increasing values of ω so that a large frequency range would have to be considered, then the real component will have to be used after all, with consideration of the following facts. Instead of Equation [14], we can also write

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi}\int_0^\infty \frac{\text{Re } F(j\omega)}{\omega} \sin \omega t d\omega = A(0) + \frac{2}{\pi}\int_0^\infty \frac{\text{Re } F(j\omega) - A(0)}{\omega} \sin \omega t d\omega \quad [18]$$

We remark that according to Equation [7]

$$A(0)\frac{2}{\pi}\int_0^\infty \frac{\sin \omega t}{\omega} d\omega$$

represents $A(0)$ times a unit step and, consequently, a step of magnitude $A(0)$. Hence, for $t > 0$ we obtain

$$A(0) - \frac{2}{\pi}\int_0^\infty \frac{A(0)}{\omega} \sin \omega t d\omega = 0$$

We assume therefore that the quantity

$$\left[\frac{\text{Re } F(j\omega) - A(0)}{\omega}\right]_{\omega=0}$$

exists (is a real number), thus allowing the graphical evaluation of the integral.

For larger values of ω, the curves $Im\ F(j\omega)/\omega$ or Re $F(j\omega)/\omega$, temporarily denoted by $K(\omega)$, can be replaced frequently by a curve $K(\omega) = c/\omega^2$ especially if, with increasing values of ω, they approach zero relatively fast and are thus suitable for graphical evaluation.

If the graphical evaluation is broken off at a value ω_g, using the foregoing relation for $K(\omega)$, then a bound on the error which thus results in the curve $\varphi(t)/\sigma_{ST}$ can be given. As we show, the error $\Delta\varphi(t)/\sigma_{ST}$ is on the order of

$$\frac{2}{\pi}\omega_g a_g \quad [19]$$

or less for $a_g = K(\omega_g)$.

If we write

$$K(\omega) = \frac{c}{\omega^2} = a_g\frac{\omega_g^2}{\omega^2}$$

then the component $\varphi_r(t)$ in $\varphi(t)$, which corresponds to the remainder of the curve, becomes, according to Equations [14] and [15]

$$\frac{\varphi_r(t)}{\sigma_{ST}} = \frac{2}{\pi}\int_{\omega_g}^{\infty} \frac{a_\varphi \omega_\varphi^2}{\omega^2} \sin \omega t d\omega \quad \ldots\ldots [20]$$

or

$$\frac{\varphi_r(t)}{\sigma_{ST}} = \frac{2}{\pi}\int_{\omega_g}^{\infty} \frac{a_\varphi \omega_\varphi^2}{\omega^2} \cos \omega t \, d\omega \quad \ldots\ldots [21]$$

If t in Equation [21] is set equal to zero, then the integral can be evaluated. Since $(\cos \omega t)_{t=0} = 1.0$ for all values of ω, we have, corresponding to Expression [19]

$$\frac{\varphi_r(0)}{\sigma_{ST}} = \frac{2}{\pi} a_\varphi \omega_\varphi \quad \ldots\ldots [22]$$

For $t > 0$, the values yielded by the integrals of Equations [20] and [21] must become smaller, since cos ωt or sin ωt now cannot become equal to 1.0 over the entire frequency range, for any value of time. Consequently the error must remain smaller than when $t = 0$ and when the imaginary component is evaluated.

Illustrative Examples

Problem 1. Let us illustrate the graphical determination of the transient output response by means of a simple problem which also could be treated exactly by calculation. Fig. 9 shows a simple closed-loop control system consisting of two components connected in series with one another. The voltage regulation of a generator by means of an integral action controller might, for example, be the physical embodiment of this control loop. The behavior of the individual components is characterized by their transient responses to a step input and has also been plotted in Fig. 9.

We then have

$$F_1 = -\frac{1}{sT_1}$$

(Negative sign indicates that controller has a corrective effect.)

$$F_2 = \frac{1}{1 + sT_2}$$

and consequently the total transfer function of the open-loop system is

$$F_R = F_1 F_2 = -\frac{1}{sT_1(1 + sT_2)}$$

As shown in Fig. 9, the disturbance will be assumed to act either at the input σ_a, or at the output of the second component σ_b. The second case, for example, would represent a set-point change. Thus, the disturbance transfer function then becomes

$$F_{\sigma_a} = \frac{1}{1 + sT_2} \quad \text{or} \quad F_{\sigma_b} = 1.0$$

and the closed-loop transfer function is

$$F = \frac{F_\sigma}{1 - F_R}; \quad F_a = \frac{sT_1}{s^2T_1T_2 + sT_1 + 1};$$

$$F_b = \frac{sT_1(1 + sT_2)}{s^2T_1T_2 + sT_1 + 1}$$

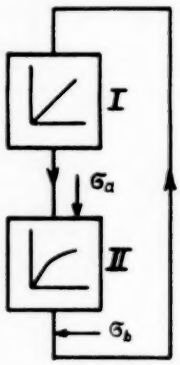


Fig. 9 Block-Diagram of a Two-Component Control System

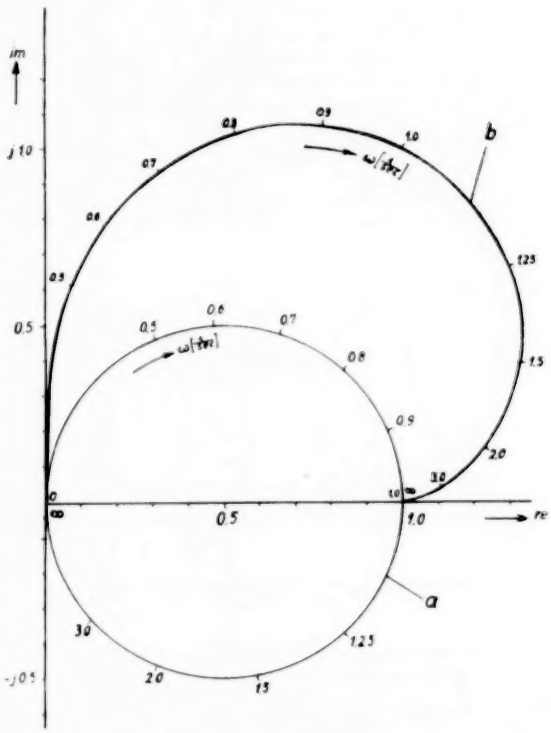


Fig. 10 Closed-Loop Frequency Response of Control System of Fig. 9

Fig. 10 shows the resulting closed-loop transfer function $F(j\omega)$ plotted in the complex plane; Fig. 11 the real and imaginary components of F_a; and Fig. 12 the real and imaginary components of F_b; all for $T_1 = T_2 = 1$ sec.

From Figs. 11 and 12, we can see immediately that, for case (a), the real component should be used; while, for case (b), the imaginary component is more suitable for the evaluation. With increasing ω, the respective curves approach zero rapidly. A check for the correctness of the calculated curves results from the fact that, for $t = 0$, the values

$$\frac{2}{\pi}\int_0^{\infty} \frac{Im\, F_a(j\omega)}{\omega} d\omega = 0$$

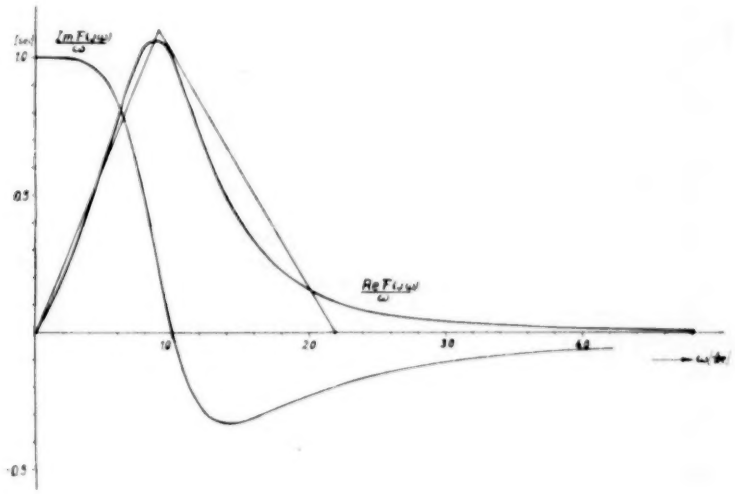


FIG. 11 Re $F(j\omega)/\omega$ AND $Im\ F(j\omega)/\omega$ FOR CURVE IN FIG. 10(a)

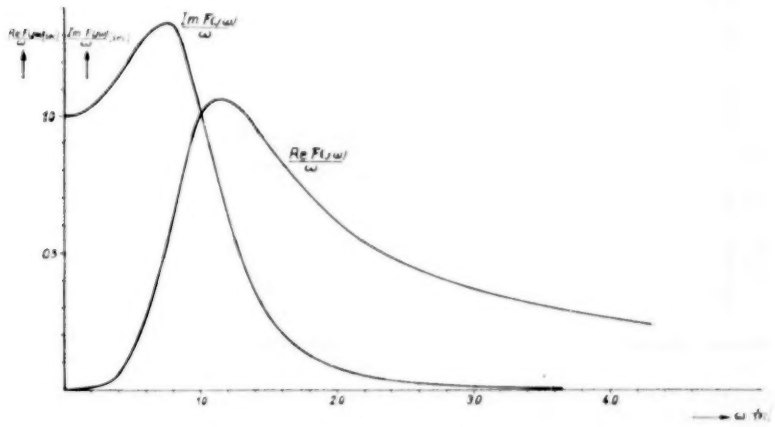


FIG. 12 Re $F(j\omega)/\omega$ AND $Im\ F(j\omega)/\omega$ FOR CURVE IN FIG. 10(b)

$$\frac{2}{\pi}\int_0^\infty \frac{Im\ F_b(j\omega)}{\omega}\,d\omega = 1.0$$

must correspond to the initial values of the curve $\varphi(t)/\sigma_{ST}$.

Fig. 13, first of all, shows the curve $(\varphi(t)/\sigma_{ST})_a$, which in this example can still be calculated quite simply. Several points marked ×, determined by means of the graphical construction indicated in Fig. 7, have been marked on the graph. In addition, Fig. 14 shows the curves

$$\frac{\text{Re } F(j\omega)}{\omega}\sin\omega t_k$$

which are necessary for the determination of the foregoing values and which have been evaluated with a planimeter.

If we approximate the curve Re $F(j\omega)/\omega$ by the triangle also shown in Fig. 11, and if we then calculate the curve $\varphi(t)/\sigma_{ST}$ according to Equation [17], a task done much more rapidly than the graphical evaluation, then the values marked ⊗ also plotted in Fig. 13 are obtained. Except for the period immediately following the beginning of the transient, they do not differ appreciably from the exact curve. Consequently, if only the approximate shape of the output transient response is of interest, then the curves Re $F(j\omega)/\omega$ or $Im\ F(j\omega)/\omega$ can be replaced by a few straight lines, as is done here, and the evaluation carried out by calculation. In Fig. 13, the calculated curve

$$\left(\frac{\varphi(t)}{\sigma_{ST}}\right)_b$$

also has been plotted from a few graphically determined values marked ×, obtained by replacement of the curve $Im\ F(j\omega)/\omega$ by a set of rectangles with width $\Delta\omega = 0.2$ (1/sec.), and subsequent summation of the term in Equation [16], given by

$$\frac{2}{\pi}\sum_{0}^{\frac{2-0}{0.2}} \frac{Im\ F(j\omega_\nu)}{\omega_\nu}\cos\omega_\nu t\Delta\omega$$

Problem 2. The control loop, shown in Fig. 15 (6), consists of three components, with component II having a constant dead time. A practical example of this control loop could, for example, be a warm-air room-heating system. Component I is equivalent to a proportional controller which regulates the heating of the warm air, while II represents the pipe conduit through which the warm air is carried to the heated room III. In this case, an exact calculated solution is not possible, since the characteristic equation be-

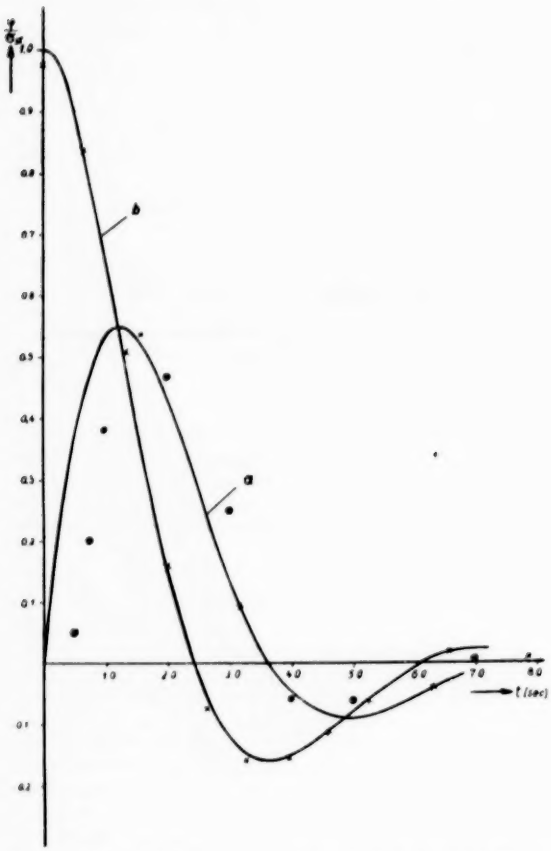


FIG. 13 TRANSIENT RESPONSE $\varphi(t)/\sigma_{ST}$ FOR SYSTEM IN FIG. 9

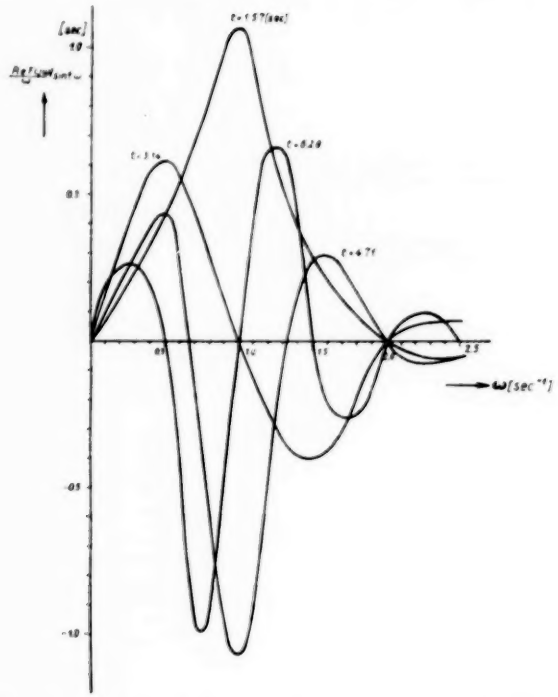


FIG. 14 AUXILIARY CURVES FOR DETERMINATION OF TRANSIENT RESPONSE IN FIG. 13

comes transcendental. Therefore, it becomes necessary to utilize approximating methods. We have

$$F_1 = -k$$

(Negative sign indicates that the controller has a corrective effect.)

$$F_2 = e^{-sT_2}$$

$$F_3 = \frac{1}{1 + sT_3}$$

$$F_R = F_1F_2F_3 = -\frac{k}{(1 + sT_3)e^{sT_2}}$$

As indicated in Fig. 15, the disturbance will be assumed acting at the input of the third σ_a, or at the input of the second component σ_b. The second case actually can be reduced to the first one, if the beginning of the output transient response is merely shifted by the time T_2, since the disturbance does not make itself felt prior to the time $t = T_2$.

In order to show, however, that the graphical method still leads to useful results even when the controlled variable assumes a constant value over a certain period of time, both cases will be treated anyway. Now

$$F_{\sigma_a} = \frac{1}{1 + sT_3}$$

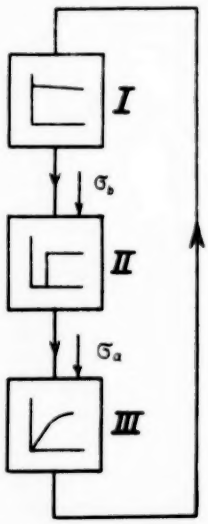


FIG. 15 BLOCK DIAGRAM OF A THREE-COMPONENT CONTROL SYSTEM

$$F_{\sigma_b} = \frac{e^{-sT_2}}{1 + sT_3}$$

and therefore

$$F_a = \frac{F_\sigma}{1 - F_R} = \frac{e^{sT_2}}{(1 + sT_3)e^{sT_2} + k}$$

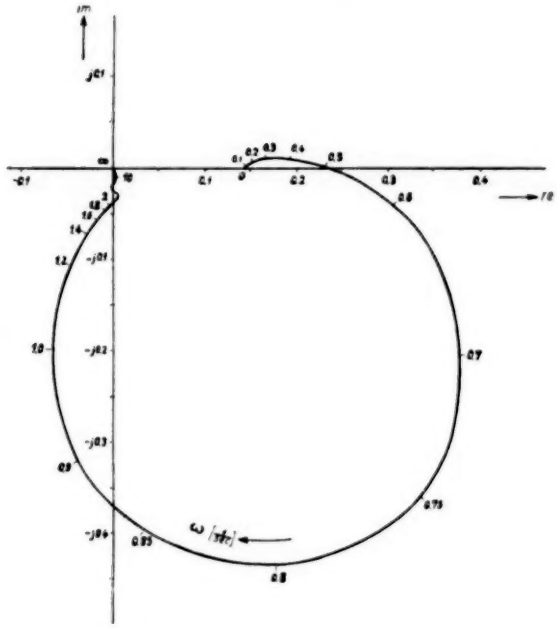


FIG. 16(a) CLOSED-LOOP FREQUENCY RESPONSE OF CONTROL SYSTEM OF FIG. 15 FOR CASE a DISTURBANCE

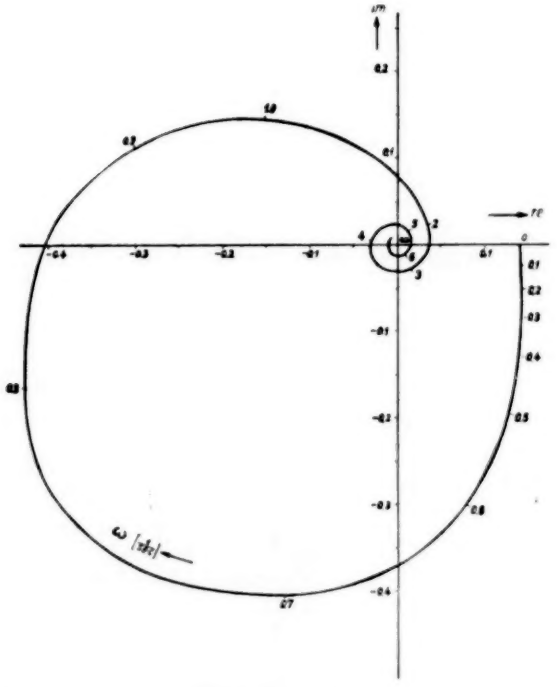


FIG. 16(b) CLOSED-LOOP FREQUENCY RESPONSE OF CONTROL SYSTEM OF FIG. 15 FOR CASE b DISTURBANCE

$$F_b = \frac{1}{(1 + sT_2)e^{sT_3} + k}$$

We will assume that $T_2 = 2$ sec, and $T_3 = 10$ sec.

The gain constant k of the open-loop system, determined by the regulator I in Fig. 15, should be selected as large as possible, so that the steady-state error (resulting from pure proportional control) will be as small as possible [$1/(1 + k)$ as compared to 1.0 without control action], and so that the error in the controlled variable will be corrected as fast as possible. If, however, the value for k is too large, then the control action becomes unstable. Since the denominator of F (for $s = j\omega$) must always be greater than zero, the limiting value $k = 8.5$ which must never be exceeded can be calculated (6). The following analysis will be carried on with $k = 6.0$.

Since, in the present problem, $F(0) = 1/(1 + k) \neq 0$, the curve [Re $F(j\omega) - A(0)]/\omega$ must be employed for the evaluation according to Equation [16], if the real component is to be used at all. Figs. 16(*a*) and 16(*b*) show the closed-loop frequency response for the two cases, while Figs. 17 and 18 show $Im\ F(j\omega)/\omega$ and [Re $F(j\omega) - A(0)]/\omega$. In both cases, the imaginary component is preferable for the evaluation.

Fig. 19 shows the evaluation for both cases, made by approximating $Im\ F(j\omega)/\omega$ by a set of rectangles of width $\Delta\omega = 0.2$ (1/sec), and by the summation

$$\frac{2}{\pi} \sum_{0}^{15} \frac{Im\ F(j\omega_\nu)}{\omega_\nu} \cos \omega_\nu t\ \Delta\omega$$

The graphs are broken off at $\omega_g = 3$(1/sec) and since, beginning with this value, the curves can be approximated by c/ω^2, the initial value for $t = 0$ has been corrected according to Equation [22]. It can be seen from the output transient response that, for case (*a*), the disturbance begins to have an effect at $t = 0$, while for case (*b*), the disturbance is not felt until $t = 2$ sec owing to the dead time of component II. Otherwise, however, the transients *a* and *b* are practically identical.

Problem 3. A more complicated control configuration, shown in the block diagram of Fig. 20, will be treated as the third problem. It also does not permit an exact calculated solution.

The following transfer functions apply to the individual components.

$$F_1 = -\frac{1 + ksT_1}{sT_1}$$

(Negative sign indicates the corrective effect of the controller.)

$$F_2 = e^{-sT_{2b}} e^{-\rho_2 \frac{sT_{2a}}{1 + sT_{2a}}}$$

$$F_3 = e^{-sT_{3b}} e^{-\rho_3 \frac{sT_{3a}}{1 + sT_{3a}}}$$

$$F_4 = e^{-sT_{4b}} e^{-\rho_4 \frac{sT_{4a}}{1 + sT_{4a}}}$$

The first component corresponds to a proportional-plus-integral-action controller. The behavior which has been assumed for the components II to IV can, for example, be found in heating systems with moving fluids, as they have been treated by Profos (7) and Takahashi (8). To clarify the meaning of the individual constants appearing in the equations, an example will be calculated in detail (7).

As shown in Fig. 21, steam is flowing through a system of pipes. The outlet temperature is to be controlled by injecting water at the inlet. The individual pipes shall be completely insulated against external heat losses. The thickness of the pipe

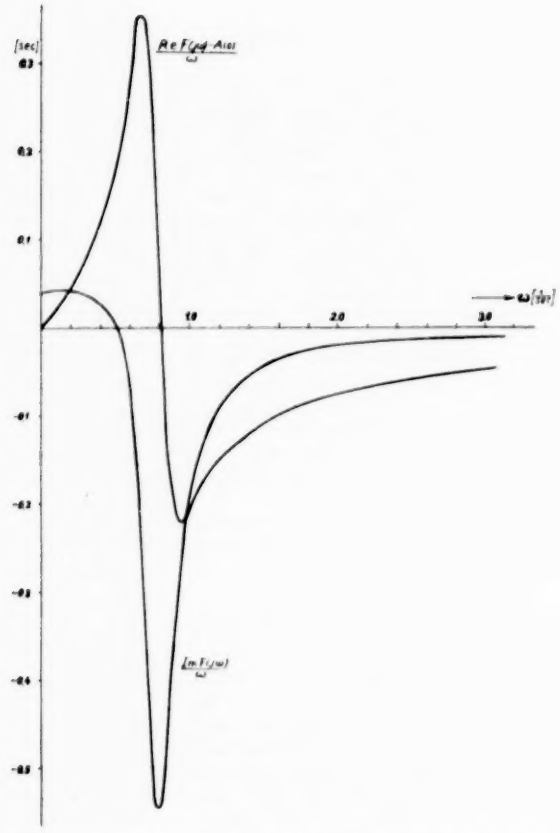


FIG. 17 Re $F(j\omega)/\omega$ AND $Im\ F(j\omega)/\omega$ FOR CURVE IN FIG. 16(a)

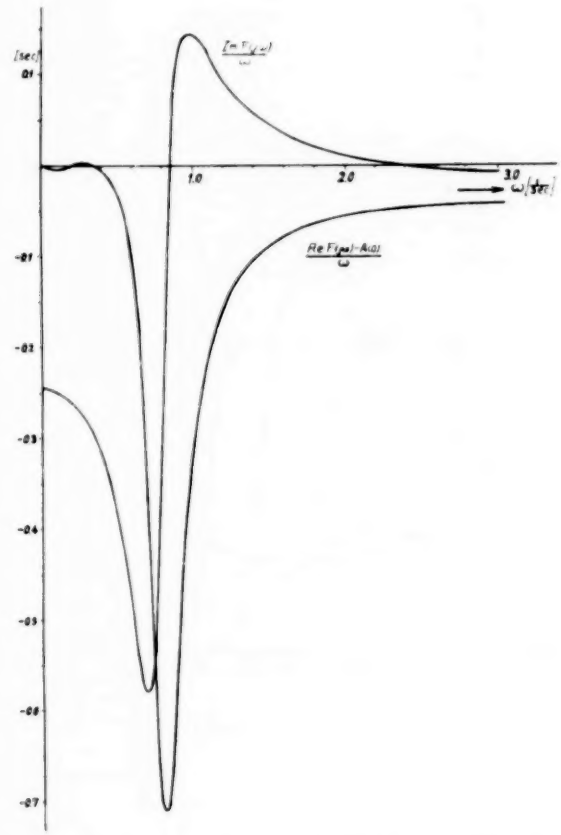


FIG. 18 $[\mathrm{Re}\ F(j\omega) - A(0)]/\omega$ AND $Im\ F(j\omega)/\omega$ FOR CURVE IN FIG. 16(b)

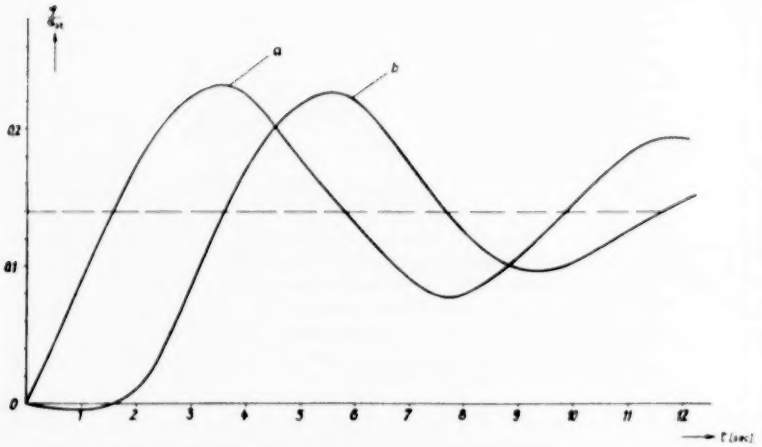


FIG. 19 TRANSIENT RESPONSE $\varphi(t)/\varphi_{81}$ FOR SYSTEM IN FIG. 15

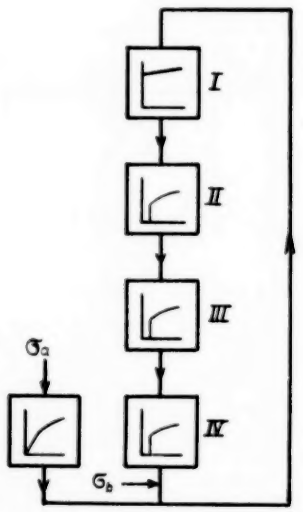


FIG. 20 BLOCK DIAGRAM OF FOUR-COMPONENT TEMPERATURE-CONTROL SYSTEM

walls, however, will not be neglected, and consequently their heat capacity must be taken into account. On the other hand, the thickness of the wall shall be considered small enough so that the wall temperature along a plane perpendicular to the direction of the flow can be assumed constant. The conduction of heat in the direction of flow will be neglected. If the entire system of pipes is now replaced by one equivalent pipe, and if a small section is cut out of this pipe, as shown in Fig. 22, then the following partial differential equations can be written

$$\frac{\partial \vartheta_b}{\partial t} C_b = -\frac{\partial I}{\partial x} \qquad [23]$$

$$\frac{\partial I}{\partial x} = \frac{\vartheta_b - \vartheta_a}{R} \qquad [24]$$

$$\frac{\partial \vartheta_a}{\partial x} = \frac{\partial \vartheta_a}{\partial t}\frac{1}{v} + \frac{\partial I}{\partial x}\frac{1}{C_a v} \qquad [25]$$

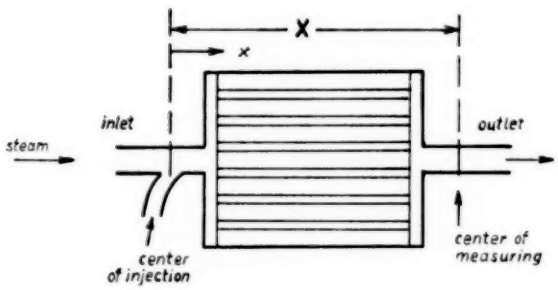


FIG. 21 SYSTEM OF PIPES CARRYING STEAM, WITH TEMPERATURE BEING CONTROLLED BY WATER INJECTION

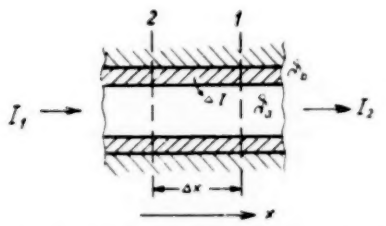


FIG. 22 SMALL SECTION OF PIPE SYSTEM SHOWN IN FIG. 21

where the symbols have the following meaning:

I = the flow of heat (quantity of heat transported by steam per unit time in x-direction)

v = velocity of steam

R = heat-transfer resistance between steam and pipe wall per unit length

ϑ = temperature (ϑ_a of steam; ϑ_b of pipe)

C = heat-capacity per unit length (C_a of steam; C_b of pipe)

By eliminating I and ϑ_b from these three equations, the partial differential equation for the steam temperature ϑ_a (from now on designated merely as ϑ) as a function of x and t is obtained

$$\frac{\partial^2 \vartheta}{\partial x \partial t} vRC_a + \frac{\partial^2 \vartheta}{\partial t^2} RC_a + \frac{\partial \vartheta}{\partial t} + \frac{\partial \vartheta}{\partial x} v\frac{RC_a}{RC_b} + \frac{\partial \vartheta}{\partial t}\frac{RC_a}{RC_b} = 0 \qquad [26]$$

We now define the following parameters:

$T_a' = RC_a = \dfrac{T_b}{\rho}$ = temperature-time constant of steam, if heated from wall of pipe

$T_a = RC_b$ = temperature-time constant of pipe wall, if heated from steam

$T_b = \dfrac{X}{v}$ = time required by steam to flow through system

$\rho = \dfrac{T_b}{T_a'}$ = ratio of time required by steam to flow through system to temperature-time constant of steam

Using these parameters, we get

$$\frac{\partial^2 \vartheta}{\partial x \partial t} XT_a + \frac{\partial^2 \vartheta}{\partial t^2} T_b T_a + \frac{\partial \vartheta}{\partial t} \rho T_a + \frac{\partial \vartheta}{\partial x} X + \frac{\partial \vartheta}{\partial t} T_b = 0 \qquad [27]$$

Let us now assume that the temperature varies sinusoidally with time at one given point, and therefore everywhere in the pipe system. The amplitude and phase, of course, will vary from point to point; but the frequency will be the same everywhere. Under this assumption, we can substitute the operator s for $\partial/\partial t$, and we obtain

$$\frac{d\mathbf{\Theta}}{dx} XsT_a + \mathbf{\Theta}s^2 T_a T_b + \mathbf{\Theta}\rho sT_a + \frac{d\mathbf{\Theta}}{dx} X + \mathbf{\Theta}sT_b = 0 \qquad [28]$$

where $\mathbf{\Theta}$ is the Laplace transform of ϑ. The partial differential equation has thus been transformed into an ordinary one (in the variables $\mathbf{\Theta}$ and x), and therefore can be solved. From the characteristic equation

$$w\mathbf{X}(sT_a + 1) + s^2 T_a T_b + s\rho T_a + sT_b = 0$$

we obtain the root

$$w = -\frac{sT_b(sT_a + 1) + s\rho T_a}{\mathbf{X}(sT_a + 1)} = -\frac{sT_b}{\mathbf{X}} - \frac{s\rho T_a}{\mathbf{X}(sT_a + 1)}$$

so that the solution of the differential equation becomes

$$\mathbf{\Theta} = Ce^{-\frac{x}{\mathbf{X}}\left(sT_b + \frac{s\rho T_a}{sT_a + 1}\right)} \qquad [29]$$

The integration constant C can be determined if, for example, the oscillation of the temperature at the beginning of the pipe ($x = 0$) is given or assumed. The ratio of the Laplace transform $\mathbf{\Theta}_x$ of the oscillation at any arbitrary location of the pipe to that at the inlet is

$$\frac{\mathbf{\Theta}_x}{\mathbf{\Theta}_0} = e^{-\frac{x}{\mathbf{X}}\left(sT_b + \frac{s\rho T_a}{sT_a + 1}\right)} \qquad [30]$$

The ratio of the Laplace transform $\mathbf{\Theta}_\mathbf{X}$ of the oscillation at the end of the conduit ($x = \mathbf{X}$) to that at the beginning ($x = 0$) becomes the transfer funct on and is given by

$$F = \frac{\mathbf{\Theta}_\mathbf{X}}{\mathbf{\Theta}_0} = e^{-sT_b} e^{-\rho\frac{sT_a}{1 + sT_a}} \qquad [31]$$

as already stated for the components II to IV.

For a given practical case, the following constants have been determined for the three components:

$T_{2a} = 16$ (sec)	$T_{2b} = 0.66$ (sec)	$\rho_2 = 0.37$
$T_{3a} = 8.4$ (sec)	$T_{3b} = 5$ (sec)	$\rho_3 = 9.5$
$T_{4a} = 17.7$ (sec)	$T_{4b} = 0.5$ (sec)	$\rho_4 = 0.5$

As shown in Fig. 20, two cases will again be assumed for the disturbance. Since, for case (*a*), a time-delay component has been introduced, any disturbance (for example, a disturbance in the heat supply to the entire pipe system, if we are dealing with a superheater) will influence the controlled variable only gradually. In case (*b*), the controlled variable is changed in a step-like manner, which is somewhat identical in effect to the shifting of the set point. Corresponding to both cases the disturbance transfer function becomes

$$F_{\sigma_a} = \frac{1}{1 + sT_\sigma}, \quad F_{\sigma_b} = 1.0$$

and hence the closed-loop transfer function is given by

$$F_a = \frac{F_{\sigma_a}}{1 - F_R} = \frac{F_{\sigma_a}}{1 - F_1F_2F_3F_4}$$

$$= \frac{\dfrac{1}{1 + sT_\sigma}}{1 + \dfrac{1 + ksT_1}{sT_1} \prod_{i=2}^{4} e^{-sT_{ib}} e^{-\rho_i \frac{sT_{ia}}{1 + sT_{ia}}}}$$

$$F_b = \frac{1}{1 + \dfrac{1 + ksT_1}{sT_1} \prod_{i=2}^{4} e^{-sT_{ib}} e^{-\rho_i \frac{sT_{ia}}{1 + sT_{ia}}}}$$

Here T_σ will be assumed to be 50 sec.

The constants of the controller (k and T_1) have to be matched to the entire system in such a manner that the optimum control action will result. This generally means that the deviation of the controlled variable from the set point should disappear as rapidly as possible. The determination of these constants will not be discussed here in detail. It will only be shown which value for the gain factor k represents the maximum permissible value from the standpoint of stability.

If the control response is to be a stable one, then

$$F = \frac{N(s)}{G(s)}$$

for $s = j\omega$ (see Equation [5]) must not have a pole; i.e., the denominator may not become zero. This requirement determines the limiting value of k for any given value of T_1 by means of the relation $G(s) = 0$. At first we will set $T_1 = \infty$, which represents pure proportional control, since the integral effect vanishes. Hence, the denominator $G(s)$ becomes

$$G(s) = 1 + k_g \prod_{i=2}^{4} e^{-sT_{ib}} e^{-\rho_i \frac{sT_{ia}}{1 + sT_{ia}}} = 0$$

so that

$$k_g \prod_{i=2}^{4} e^{-sT_{ib}} e^{-\rho_i \frac{sT_{ia}}{1 + sT_{ia}}} = k_g e^{-b(\omega)} e^{-j\beta(\omega)} = -1.0$$

If this requirement is to be fulfilled, then we must have $\beta(\omega) = \pi$ or 3π, 5π, and so on, because only then can the modulus of the vector Π assume negative real values. Of practical interest, however, is only the value for $\beta(\omega) = \pi$ corresponding to the predominant oscillation. Therefore, we must first determine this frequency (done best graphically), at which the vector Π is shifted by π; and then we must calculate the vector's modulus at this frequency, which will be denoted by B. Then $k_g = 1/B$.

In our problem we obtain $\beta(\omega) = \pi$ for $\omega = 0.0345$ (1/sec) with $B = 0.4$. Therefore

$$k_g = \frac{1}{0.4} = 2.5$$

For proportional control alone, k therefore must be less than 2.5. The integral action of the controller ($T_1 < \infty$), however, has a detrimental effect on the stability; and k must therefore be made substantially smaller than 2.5. In this problem, the values $k = 1.0$ and $T_1 = 100$ (sec) were used, which as the results show lead to a satisfactory control response. The constants $k = 1.0$ and $T_1 = 100$ (sec) are thus also fixed.

Fig. 23 shows the frequency response

$$F_{24} = F_2F_3F_4 = \prod_{i=2}^{4} e^{-sT_{ib}} e^{-\rho_i \frac{sT_{ia}}{1 + sT_{ia}}}$$

using the stated numerical values; Fig. 24 shows $Im\ F_{24}(j\omega)/\omega$.

Evaluation of this curve according to

$$\frac{2}{\pi} \int_0^\infty \frac{Im\ F_{24}(j\omega)}{\omega} \cos \omega t d\omega$$

by graphical integration [subdivision into sectors of width $\Delta\omega = 0.005$ (1/sec) and summation of

$$\frac{2}{\pi} \sum_0^{14} \frac{Im\ F_{24}(j\omega_\nu)}{\omega_\nu} \cos \omega_\nu t \Delta\omega]$$

yields the transient response of this section of the control loop, Fig. 25. If the curve in Fig. 24 is approximated very roughly by a straight line, also shown in the figure, and if the transient response is calculated by means of Equation [17], then we obtain the points also plotted in Fig. 25. They differ only insignificantly from the more exact curve, which was calculated with an appreciably greater amount of effort.

In Figs. 26 to 28, the frequency response, Re $F(j\omega)/\omega$ and $Im\ F(j\omega)/\omega$ for the entire closed-loop system have been drawn for both cases (*a*) and (*b*). It can be seen immediately that, for case (*a*) (disturbance gradually increasing from zero), the real component should be used for the graphical evaluation; while, for case (*b*) (step function disturbance), the imaginary component is more appropriate. Fig. 29 shows the output transient response for both cases, again evaluated by means of a summation according to Equations [16 *a*, *b*] with $\Delta\omega = 0.005$ (1/sec) and considering the curves only in the region between $\omega_a = 0$ and $\omega_b = 0.07$ (1/sec). Since the response is very well damped, the selection of the control constants (k and T_1) is justified as appropriate.

Here it can also be demonstrated how the troublesome and time-consuming graphical integration can be avoided if it is only necessary to determine the approximate shape of the output response curve. If the curve Re $F(j\omega)/\omega$ is approximated by the rectangle drawn in Fig. 27, then we obtain, according to Equation [17]

$$\frac{2}{\pi} \int_0^\infty \frac{\text{Re } F(j\omega)}{\omega} \sin \omega t d\omega \approx \frac{2}{\pi} \int_{\omega_1}^{\omega_2} a \sin \omega t d\omega$$

$$= \frac{2}{\pi} \frac{a}{t} (\cos \omega_1 t - \cos \omega_2 t) = \frac{\varphi(t)}{\sigma_{ST}}$$

with $a = 51$ (sec); $\omega_1 = 0.0025$ (1/sec); $\omega_2 = 0.03$ (1/sec) and the shape of the curve can be found readily.

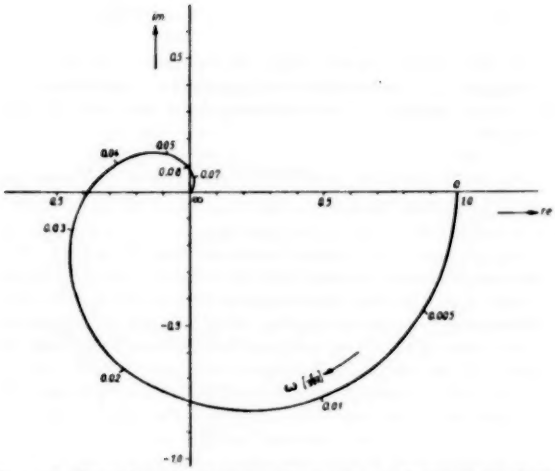


FIG. 23 OPEN-LOOP FREQUENCY RESPONSE OF COMPONENTS II TO IV OF FIG. 20

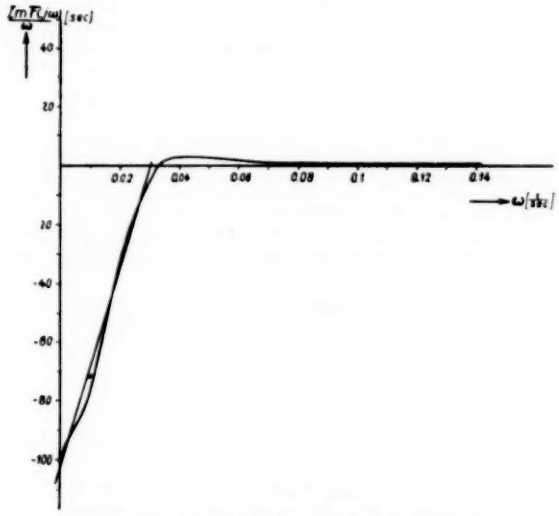


FIG. 24 *Im* $F(j\omega)/\omega$ FOR CURVE IN FIG. 23

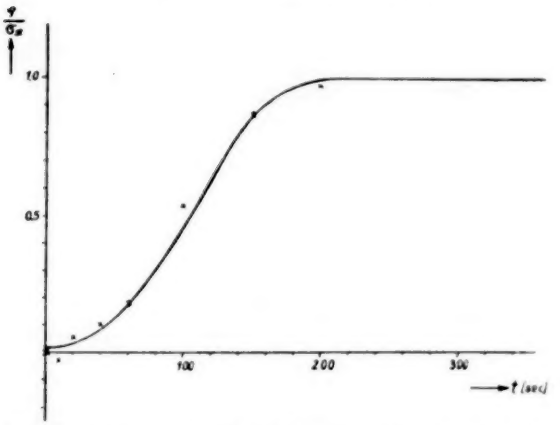


FIG. 25 OPEN-LOOP TRANSIENT RESPONSE AS DETERMINED FROM CURVE IN FIG. 23

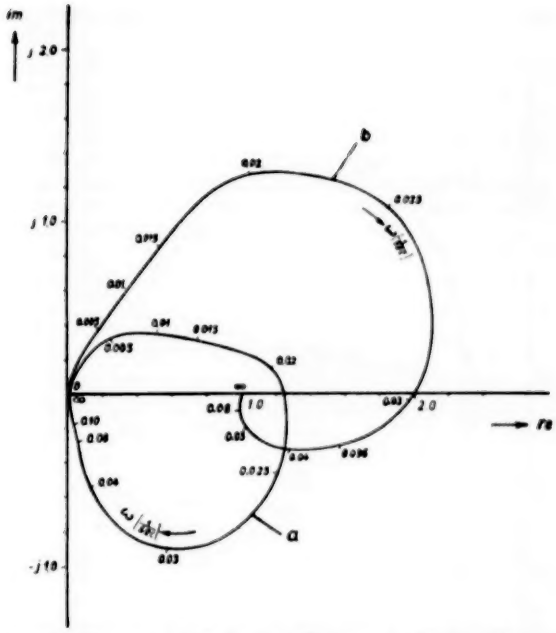


FIG. 26 CLOSED-LOOP FREQUENCY RESPONSE OF CONTROL-SYSTEM IN FIG. 20

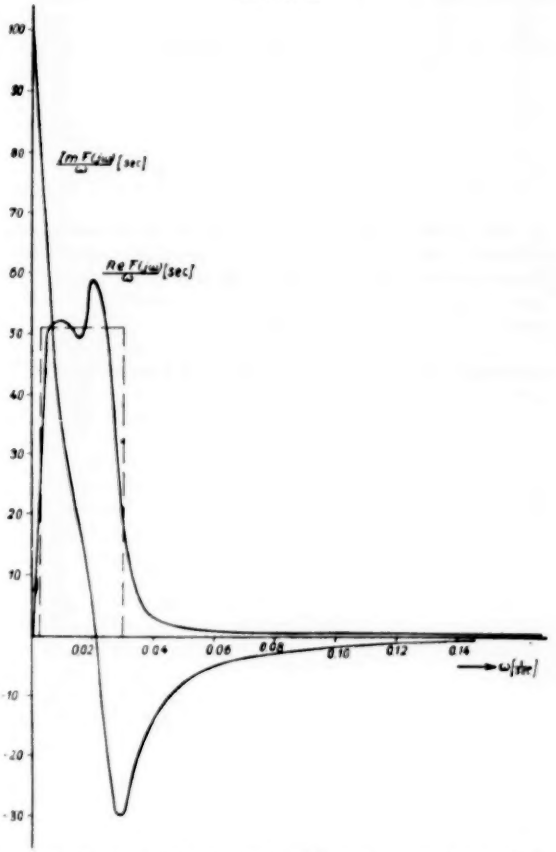


FIG. 27 Re $F(j\omega)/\omega$ AND *Im* $F(j\omega)/\omega$ FOR CURVE IN FIG. 20(*a*)

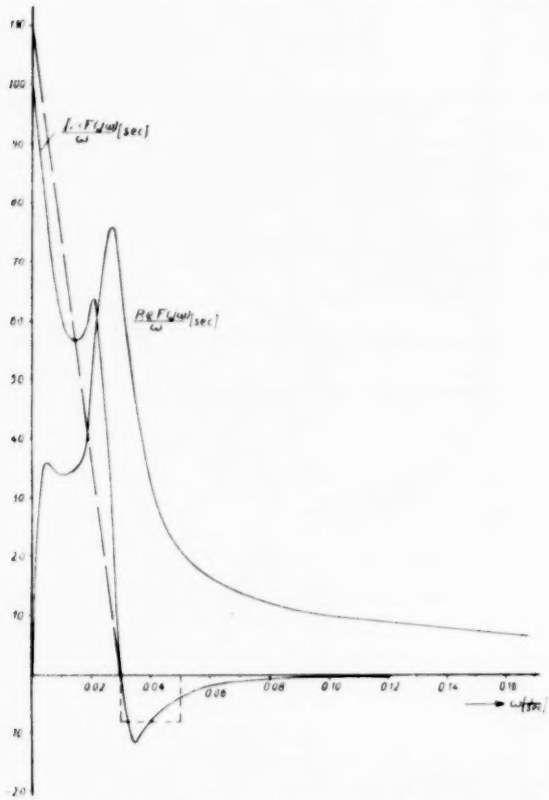


FIG. 28 Re $F(j\omega)/\omega$ AND $Im\ F(j\omega)/\omega$ FOR CURVE IN FIG. 20(*b*)

A few values, calculated in this manner have been plotted (✕) in Fig. 29(*a*). If curve $Im\ F(j\omega)/\omega$ is approximated by the triangle and rectangle shown in Fig. 28, and if the integral

$$\int \frac{Im\ F(j\omega)}{\omega} \cos \omega t d\omega$$

is calculated, then the values also plotted in Fig. 29(*b*) result. It can be seen that the approximation of the curves Re $F(j\omega)/\omega$ or $Im\ F(j\omega)/\omega$ by a few straight sections yields an approximation of the output transient response which is entirely sufficient for many purposes.

SUMMARY

A step-function disturbance can be represented by a constant value and an infinite number of undamped sinusoidal oscillations, as long as linear characteristics are assumed for the system. The closed-loop transfer function $F(j\omega)$ is defined as the ratio of the complex quantity denoting the sinusoidal oscillation of the controlled variable to the complex quantity denoting the oscillation of the input disturbance, for frequencies between zero and infinity. If the closed-loop frequency response $F(j\omega)$ is known, then the effect which each individual oscillation of the step disturbance has on the controlled variable can be found. The controlled variable, in turn, now also appears as the sum of an infinite number of undamped sinusoidal oscillations.

A summation of all these individual oscillations yields the response of the controlled variable as a function of time; i.e., the so-called output transient response. In carrying out its graphical evaluation, we can limit ourselves usually to a very narrow frequency range, since the effect of oscillations with frequencies outside of this range is very small and can be neglected. The graphical method can now be based directly on the relation

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \int_0^\infty A(\omega) \frac{\sin [\omega t + \alpha(\omega)]}{\omega} d\omega \quad [9]$$

where $A(\omega)$ and $\alpha(\omega)$ are assumed to be constant within small segments $\Delta\omega$, and where only the frequency range which affects the transient is being considered. The integral thus turns into a sum, which can be determined easily.

For an approximate evaluation, the relations

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi} \int_0^\infty \frac{\text{Re } F(j\omega)}{\omega} \sin \omega t d\omega \quad [14]$$

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi} \int_0^\infty \frac{Im\ F(j\omega)}{\omega} \cos \omega t d\omega \quad [15]$$

which can be derived easily from Equation [9] appear to be even more appropriate. For each individual problem it must be de-

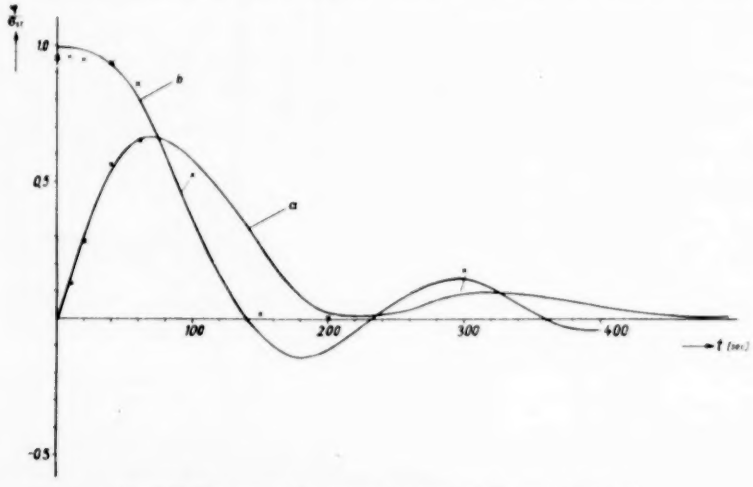


FIG. 29 TRANSIENT RESPONSE $\varphi(t)/\sigma_{ST}$ FOR SYSTEM IN FIG. 20

cided whether it is more practical to use Re $F(j\omega)$ or $Im\ F(j\omega)$ for the evaluation. The graphical determination of the integrals can be carried out in various ways, and mechanical calculating devices can be employed to advantage. A rough approximation of the curves Re $F(j\omega)/\omega$ or $Im\ F(j\omega)/\omega$ by a few straight lines frequently yields useful information about the shape of the output transient response.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Mr. D. W. Pessen[3] for revision of the paper's English translation and to Mr. Mark Leum[4] and Dr. Rufus Oldenburger[5] for their valuable suggestions. He also wishes to thank his co-workers, Mr. Ernst and Mr. Poensgen, who assisted him in translating the paper into English and also in calculating the curves.

BIBLIOGRAPHY

1 "Operatorenrechnung," by K. W. Wagner, J. A. Barth-Verlag, Leipzig, Germany, 1940.

2 "Graphische Verfahren zur Untersuchung von Regel und Steuervorgängen," by K. Mueller, Dissertation, Technische Hochschule, Stuttgart, Germany, 1953.

3 "Lösungsverfahren der Laplace-Transformation für Ausgleichsvorgänge in linearen Netzen, angewandt auf selbsttätige Regelungen," by E. Grünwald, *Archiv für Elektrotechnik*, vol. 35, 1941, p. 379.

4 "Regelungstheorie," by J. Hänny, A. G. Leemann & Company, Zürich, Switzerland.

5 "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948.

6 "Die selbsttätige Regelung," by A. Leonhard, Springer-Verlag, Berlin, Göttingen, Heidelberg, Germany, 1949.

7 "Vektorielle Regeltheorie," by P. Profos, A. G. Leemann & Company, Zürich, Switzerland, 1944.

8 "Regeltechnische Eigenschaften von Gleich- und Gegenstromwärmeaustauschern," by Y. Takahashi, *Regelungstechnik*, vol. 1, 1953, p. 32.

"Transfer Function Analysis Heat Exchange Processes," by Y. Takahashi, *Automatic and Manual Control*, Butterworth Publications, Ltd., London, England, 1951.

Discussion

A. TUSTIN.[5] The writer would call attention to a particular form of mechanical harmonic analyzer that was devised and constructed at the University of Birmingham for the purpose of interconverting time and frequency response functions, on principles similar to those described in this paper.

This instrument has been described in detail by J. G. Henderson.[6] The instrument gives the Fourier coefficients, not for the curve analyzed, but for its derivative. Thus it is directly adapted to solve the inverse problem to that chiefly discussed in the paper. Given a response to a unit step function, obtained experimentally, it will plot the frequency response locus. If $A(t)$ is the response to unit function the relationships that give the vector components $R(\omega)$ and $I(\omega)$ of the response vector at frequency ω are

$$R(\omega) = \int_0^\infty \dot{A}(t) \cos \omega t dt$$

$$I(\omega) = -\int_0^\infty \dot{A}(t) \sin \omega t dt$$

where $\dot{A}(t)$ is the derivative of $A(t)$.

The principle of the device is shown in Fig. 30 of this discussion. A flat plate is free to move in either or both of two directions at right angles. A friction wheel rests on it and drives it. When the curve $A(t)$ to be analyzed is tracked it causes the wheel, through stretched wire connections, to be turned about a vertical axis by an amount ωt, and at the same time to be turned about its horizontal axis by $\delta A(t)$, the change of the height or ordinate of the curve $A(t)$, during the time δt.

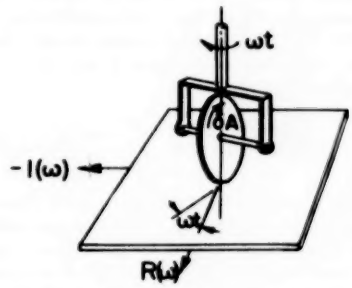


FIG. 30 PRINCIPLE OF HARMONIC ANALYZER USING WHEEL AND RESOLVER PLATE

This results in the plate being displaced by an amount proportional to $\delta A(b)$ in a direction given by the angle ωt. The components of the displacement when the complete response curve has been traversed are therefore proportional to

$$\int_0^\infty \frac{\delta A}{\delta t} \cos \omega t \delta t = R(\omega)$$

and

$$\int_0^\infty \frac{\delta A}{\delta t} \sin \omega t \delta t = -I(\omega)$$

The displacement of the plate for a complete traversal of the curve, from $t = 0$ until no further changes δA occur, is thus the vector of the response-vector locus for the particular frequency ω. The various values of ω are obtained by settings of a variable-speed gear provided between the displacement (t) of the table carrying the curve, and the rotation of the friction wheel (ωt).

This form of harmonic analyzer also may be used to carry out the calculations required in the problems considered by the author of the paper, that all depend upon integrations of similar form.

C. H. DAWSON.[7] The author has contributed an approach to the problem of predicting time response in those linear systems where analytical treatment is impossible or very cumbersome owing to the frequency response being known only in graphical form, the presence of dead lag in the system, or the characteristic equation being of high order. In this problem selection of a single type of time input for consideration obviously is required, the unit step and unit pulse being particularly convenient. When using the unit pulse, which is the time derivative of the unit step, the output is given by

$$c(t) = \frac{2}{\pi} \int_0^\infty \text{Re } F(j\omega) \cos \omega t d\omega \quad \ldots\ldots [32]$$

or

$$c(t) = -\frac{2}{\pi} \int_0^\infty Im\ F(j\omega) \sin \omega t d\omega \quad \ldots\ldots [33]$$

[3] Brown Instruments Division, Minneapolis-Honeywell Regulator Company, Philadelphia, Pa.

[4] Woodward Governor Company, Rockford, Ill.

[5] Professor, University of Birmingham, Birmingham, England.

[6] "Mechanical Harmonic Analyzer and Some Applications to Servo Systems," by J. G. Henderson, *Engineering*, vol. 173, 1952, Jan. 11, pp. 52–53; Jan. 18, pp. 68–70.

[7] Associate Professor of Electrical Engineering, University of Rochester, Rochester, N. Y.

which are easily obtainable from the author's Equations [14] and [15] by differentiation with respect to time. Floyd's work in reference (5) of the paper uses Equation [32], while Dawson's work[8] uses Equation [33]. Regardless of the type of input selected, the heart of the probem is the approximate integration of a function of the form $K(\omega)\sin t\omega$ or $K(\omega)\cos t\omega$. Since the functions $K(\omega)$ in all cases decrease rapidly with increasing frequency, the integration can be restricted to a finite range.

If the integral selected involves $\cos t\omega$, the error due to finite integration can be evaluated for $t = 0$ as the author has done in Equation [21]. However, the error so determined is the maximum involved only when the $\cos t\omega$ form is used since there is no reason to suppose that

$$\int_{\omega g}^{\infty} \frac{ag\ \omega g^2}{\omega^2} \sin \omega t d\omega \leq \int_{\omega g}^{\infty} \frac{ag\ \omega g^2}{\omega^2} \cos \omega t d\omega \quad \ldots [34]$$

unless Re $F(\omega)/\omega$ approaches zero more rapidly than $Im\ F(\omega)/\omega$ with increasing frequency. If Re $F(\omega)$ does approach zero more rapidly and is accordingly chosen for integration, then the error determined from Equation [21] of the paper may be considerably higher than the true error; it will, however, be on the safe or pessimistic side.

The transfer of information from the closed-loop polar plot such as Fig. 10 to the component against frequency plots of Fig. 11 can be expedited if the horizontal and vertical straight-line loci of constant Re $F(\omega)$ and $Im\ F(\omega)$ are superimposed on the polar plot. However, system design usually proceeds from consideration of the open-loop Nyquist plot or Black presentation of log magnitude against angle and it is possible to transfer directly from the open-loop data to the real and imaginary components of the closed-loop response through loci of the type developed in reference (5) of the paper and by the writer.[8] The closed-loop frequency response is not needed for this purpose.

Except in special cases where $K(\omega)$ can be approximated by straight-line segments, the sums of Equations [16a] and [16b] of the paper contain $\sin t\omega_r$ or $\cos t\omega_r$ inside the summation sign indicating that the summing of several sinusoidal terms of different frequencies for each time t is required. If $K(\omega)$ is expanded as a Fourier sine or cosine series in ω over the range 0 to Ω where $K(\Omega)$ is essentially zero, and this series is substituted in Equation [14] or [15], the resulting expression is much easier to handle. For example, suppose in Fig. 5 of the paper, Ω is taken equal to $3\omega_r$. Then, using a Fourier sine series of period 2Ω

$$\left.\begin{aligned} \frac{\text{Re}\, F(j\omega)}{\omega} &= \sum_{n=1}^{\infty} a_n \sin n \frac{\pi\omega}{\Omega} && 0 < \omega < \Omega \\ &= 0 && \Omega < \omega \end{aligned}\right\} \quad \ldots [35]$$

where a_n can be determined by a method similar to that of Fig. 7, using the defining integral

$$a_n = \frac{2}{\Omega} \int_0^{\Omega} K(\omega) \sin \frac{n\pi\omega}{\Omega} d\omega \quad \ldots [36]$$

Then from Equation [14] of the paper

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi} \int_0^{\infty} \frac{\text{Re}\, F(j\omega)}{\omega} \sin \omega t dw \quad \ldots [14]$$

$$= \frac{2}{\pi} \int_0^{\Omega} \sum_{n=1}^{\infty} a_n \sin n \frac{\pi\omega}{\Omega} \sin \omega t d\omega$$

$$= \sum_{n=1}^{\infty} \frac{2}{\pi} \int_0^{\Omega} a_n \sin n \frac{\pi\omega}{\Omega} \sin \omega t d\omega$$

$$= \sum_{n=1}^{\infty} \frac{a_n \Omega}{\pi} \left[\frac{\sin(n\pi - \Omega t)}{n\pi - \Omega t} - \frac{\sin(n\pi + \Omega t)}{n\pi + \Omega t} \right]$$

$$= 2\Omega \sin \Omega t \sum_{n=1}^{\infty} \frac{(-1)^{n+1} n a_n}{n^2\pi^2 - \Omega^2 t^2} \quad \ldots [37]$$

Note that in this equation only one sine term need be evaluated for each time t. If Ωt is then set equal to τ, the response $\phi(\tau)$ is easily determined by tabular computation on prepared forms. Detailed consideration of this derivation and its use are contained in the writer's paper.[8]

The author's comments on harmonic analysis following Equation [16b] seem to apply to an attempt to use t as the summation index of the series whose coefficients would be determined by Equation [14] and hence are concerned with a procedure essentially different from that proposed in this comment. However that may be, in using the foregoing method, the Fourier coefficients were obtained readily by graphical integration and the use of only the first four terms of the resulting series was found to give quite accurate results for the systems investigated.

S. R. Sporn.[9] The author has reviewed concisely the necessity for having available a graphical method of obtaining transient response from frequency response. However, the method developed in the paper (although simpler than some methods previously proposed) is certainly not the most powerful that can be constructed in the present state of the art.

The writer and Mr. Monroe Berriman of the Arma Corporation are collaborating on a paper which will present an approach to general synthesis. The basic method used is to apply the sampling theorem proposed by Shannon.[10] The method occurred to each of us independently, Mr. Berriman developing the basic results in his master's thesis,[11] the writer utilizing a proof of Shannon's theorem for the same results.

Professor Dawson's method, as outlined in his discussion, may be extended to be a much more powerful tool, in which case it would become identical to the writer's approach. This can be shown by deriving Dawson's Equation [37] from the sampling theorem and proof.

A statement of Shannon's theorem appears in Equation [38], herewith, while Equation [39] is developed in the proof of Equation [38]

$$\phi(t) = \sum_{-\infty}^{+\infty} \phi\left(\frac{n}{2W}\right) \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} \quad \ldots [38]$$

$$\phi\left(\frac{n}{2W}\right) = 2Wa_n \quad \ldots [39]$$

where the actual Re$[F(j\omega)]$ is approximated by a band-width-limited function having a cutoff frequency of W cps.

From Equation [38]

[8] "Approximation of Transient Response From Frequency Response Data," by C. H. Dawson, Trans. AIEE, vol. 72, 1953, AIEE Technical Paper 53-248.

[9] Development Engineer, Arma Corporation, Garden City, Long Island, N. Y.

[10] "Communication in the Presence of Noise," by C. E. Shannon, Proceedings of the IRE, vol. 37, 1949, pp. 10–21.

[11] The Determination of the Time Function of a Servomechanism From Its Frequency Response in Graphical Form," by M. Berriman, a Thesis for Master of Science in Electrical Engineering, Ohio State University, 1948.

$$\phi(t) = \sum_{-\infty}^{+0} \phi\left(\frac{n}{2W}\right) \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} + \sum_{+1}^{\infty} \phi\left(\frac{n}{2W}\right) \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

$$\phi(t) = \sum_{\infty}^{0} \phi\left(\frac{-n}{2W}\right) \frac{\sin \pi(2Wt + n)}{\pi(2Wt + n)} + \sum_{+1}^{\infty} \phi\left(\frac{n}{2W}\right) \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

But

$$\phi\left(\frac{-n}{2W}\right) = \phi\left(\frac{n}{2W}\right)$$

since $\phi(t)$ can be assumed even so that $Im\ [F(j\omega)]$ vanishes.

$$\phi(t) = \sum_{+1}^{\infty} \phi\left(\frac{n}{2W}\right) \left[\frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} - \frac{\sin \pi(2Wt + n)}{\pi(2Wt + n)}\right]$$

Expanding the sin terms and combining

$$\phi(t) = \sum_{+1}^{\infty} \phi\left(\frac{n}{2W}\right) \sin 2\pi Wt \frac{(-1)^{n+1} 2\pi n}{[(\pi n)^2 - (2\pi Wt)^2]}$$

and substituting Equation [39]

$$\phi(t) = 2W \sin 2\pi Wt \sum_{n=1}^{\infty} (-1)^{n+1} \frac{2\pi n a_n}{(\pi n)^2 - (2\pi Wt)^2}$$

which is identical to Dawson's Equation [37], $\Omega \equiv 2\pi W$.

Note that Equation [39] states that each harmonic of the frequency response uniquely determines the time function at a perticular instant of time (sample) while Equation [38] enables the time function to be computed for other times. Equation [39] is the extension of Dawson's work.

Thus the labor in computing the time function from a frequency function reduces to the empirical computation of a Fourier series (a small number of terms are usually sufficient) of sines or cosines—an extremely simple procedure as most engineers who have calculated the distortion in an electronic amplifier will admit.

For further information concerning the foregoing results, i.e., derivation, conditions of applicability, accuracy, and application to synthesis, the reader is referred to the forthcoming paper.

E. C. Grogan.[12] The author has presented a clear-cut approach to the problem of correlating the transient behavior of systems with their steady-state frequency response. He has developed techniques which are readily understood by those familiar with the principle of superposition, applied to systems with linear characteristics. This discussion will cover the significance of the expressions given in the paper for approximating the output response of systems, when forced by step input functions.

The expression used by the author for representing the response of a system to a step input function is given as[13]

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \int_0^{\infty} A(\omega) \frac{\sin[\omega t + \alpha(\omega)]}{\omega} d\omega \quad \ldots [9]$$

[12] Research and Development Engineer, Brown Instruments Division, Minneapolis-Honeywell Regulator Company, Philadelphia, Pa.

[13] Numbers of equations in this discussion correspond with those given in the paper.

where the transfer function of the system is represented by

$$F(j\omega) = A(\omega) e^{j\alpha(\omega)}$$

As a basis for understanding the development of Equation [9], it may be helpful to consider the response of a system to a repeating saw-tooth input function, illustrated in Fig. 31 of this discussion and expressed by the Fourier series

$$\frac{\sigma(t)}{\sigma_{ST}} = \frac{1}{2} + \frac{1}{\pi}\left[\sin \omega_a t + \frac{1}{2} \sin 2\omega_a t + \frac{1}{3} \sin 3\omega_a t + \ldots\right]$$

or

$$\frac{\sigma(t)}{\sigma_{ST}} = \frac{1}{2} + \frac{1}{\pi} \sum_{v=0}^{\infty} \frac{\sin \omega_v t \Delta\omega}{\omega_v}$$

where

$$\omega_v = \omega_a + v\Delta\omega; \quad \frac{\omega_a}{\Delta\omega} = 1$$

Applying the transfer function of the system to the harmonic components comprising the foregoing, the output response may be written as

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \sum_{v=0}^{\infty} A(\omega_v) \frac{\sin[\omega_v t + \alpha(\omega_v)]}{\omega_v} \Delta\omega$$

Referring to Fig. 31, it becomes evident that as $\omega_a \to 0$, the output response of the system, i.e., over its settling period, will approach its theoretical response to a step input function; since

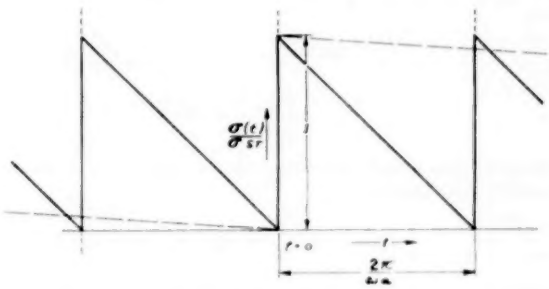


Fig. 31 Saw-Tooth Function

the decay rate of the saw-tooth input function at $0 - > t > 0 +$ is shown to approach zero as its period approaches infinity. Thus, noting that

$$\lim_{\omega_a \to 0} \sum_{v=0}^{\infty} A(\omega_v) \frac{\sin[\omega_v t + \alpha(\omega_v)]}{\omega_v} \Delta\omega = \int_0^{\infty} A(\omega) \frac{\sin[\omega t + \alpha(\omega)]}{\omega} d\omega$$

the output response of a system to a step input function may be expressed in terms of its frequency response, as indicated by Equation [9].

It is interesting to note that when graphical methods of integration are applied to Equation [9], the resulting approximate solution for output response, in many cases, will represent the response of a system to a repeating rectangular wave input function. For example, when Equation [10] is applied to a system with finite gain at zero frequency, and with frequency-response curves that may be divided into a number of sectors of equal

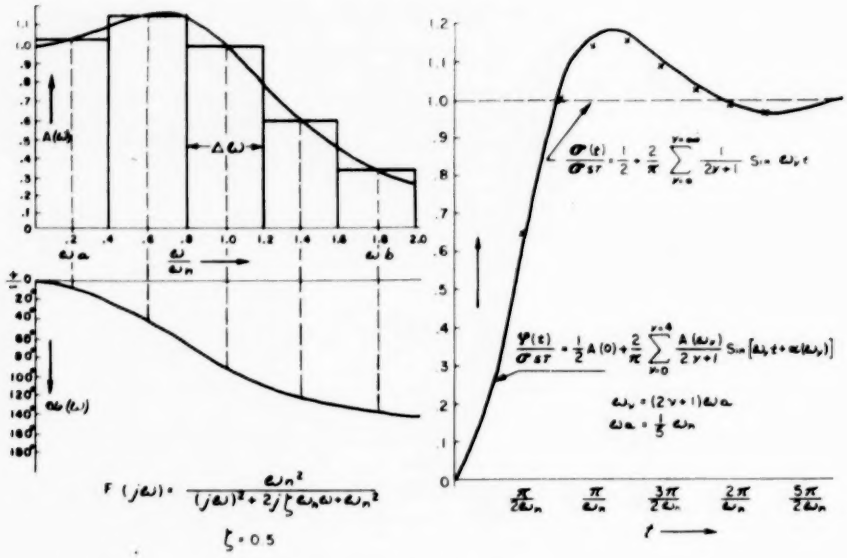


FIG. 32 RESPONSE TO REPEATED RECTANGULAR INPUT

width $\Delta\omega$, as shown in Fig. 32, herewith, Equation [10] will be of the form[14]

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{2}{\pi}\left\{A(\omega_a)\sin[\omega_a t + \alpha(\omega_a)] + \frac{A(3\omega_a)}{3}\sin[3\omega_a t + \alpha(3\omega_a)] + \frac{A(5\omega_a)}{5}\sin[5\omega_a t + \alpha(5\omega_a)] + \ldots + \left[\frac{\omega_a}{\omega_b}\right] A(\omega_b)\sin[\omega_b t + \alpha(\omega_b)]\right\}$$

Equation [10] would then represent the response of a system forced by the function

$$\frac{\sigma(t)}{\sigma_{ST}} = \frac{1}{2} + \frac{2}{\pi}\left[\sin\omega_a t + \frac{1}{3}\sin 3\omega_a t + \frac{1}{5}\sin 5\omega_a t + \ldots\right]$$

which is the Fourier-series expression for the rectangular wave illustrated in Fig. 33 of this discussion.

The system used in the example given in Fig. 32 has an underdamped second-order characteristic, with unit gain at zero frequency. Five frequency terms are used in the approximation, with a base frequency that is one fifth the natural frequency of the system. The resulting plot of output response is shown at the right of Fig. 32 and by curve *A* in Fig. 34.

When viewed over the apparent settling period of the system, this curve is shown to be almost coincident with the exact solution for system response to a step input function, which is indicated by cross marks.

The technique of correlating frequency-response data with transient phenomena, by applying the transfer function of the system to the individual harmonic terms contained in the Fourier-series expression for a square wave, is used frequently with the aid of electromechanical harmonic synthesizers. If one half the period of the base frequency is large compared with the apparent

[14] In this case, the base frequency ω_a, which is the frequency at the mean ordinate of the first sector, is one half the frequency increment $\Delta\omega$. The frequency of the sine terms used in the approximation will then be odd integer multiples of the base frequency.

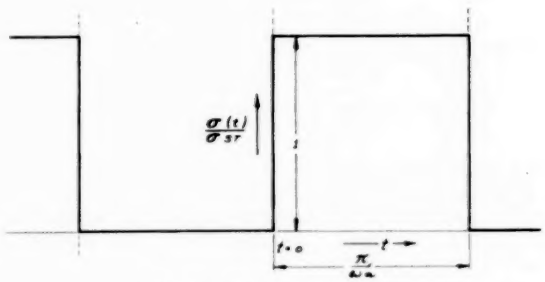


FIG. 33 RECTANGULAR WAVE

settling time of the system, a good approximation of system response to a step function should be obtained.[15]

Equation [10*a*], given in the lower-left corner of Fig. 34, is essentially the same as Equation [10] of the paper with the exception that t has been replaced by $(-t)$. A plot of this function is shown by curve *B*. The value of this function is essentially zero from $t = 0$ to $t = T_C$, the apparent settling time of the system (with the exception of the small-amplitude oscillation about the zero axis which is inherent in the approximation).

Equations [10] and [10*a*] can be transformed into expressions involving the real and imaginary components associated with the transfer function of the system. When Equation [10] is added to Equation [10*a*] the terms involving the real components drop out, resulting in Equation [16*b*].

Curve *C* shown in the upper-right corner of Fig. 34 would be the plot of this equation, which is shown to be curve *A* superimposed upon curve *B*. Similarly, when Equation [10*a*] is subtracted from Equation [10] the terms involving the imaginary components drop out, resulting in Equation [16*a*]. A plot of this function would then be curve *B* subtracted from curve *A*, shown as curve *D* in the lower-right corner of Fig. 34.

[15] Since this approximation assumes sharp cutoff properties for the system (i.e., $A(\omega) = 0$ for $\omega > \omega b$), the Gibbs phenomena in the Fourier series for a rectangular wave will have some influence on the solution for output response. The degree to which this effect is apparent will depend upon the number of terms used in the approximation.

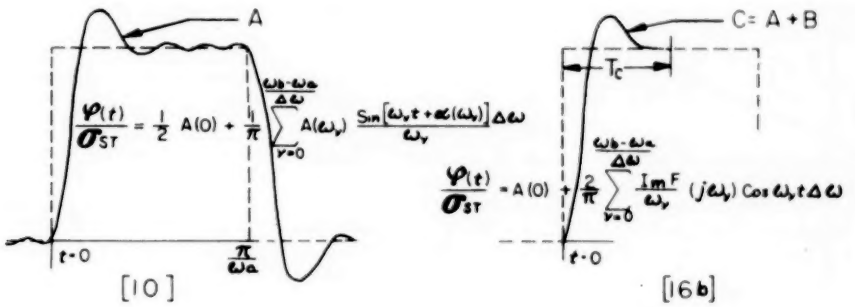
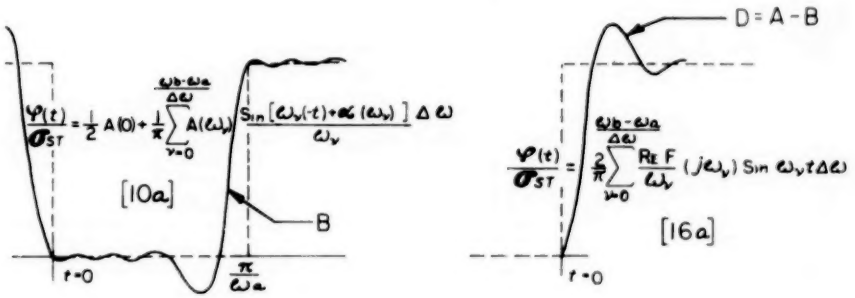

A
C = A + B
$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \sum_{\nu=0}^{\frac{\omega_b-\omega_a}{\Delta\omega}} A(\omega_\nu) \frac{\sin[\omega_\nu t + \alpha(\omega_\nu)]\Delta\omega}{\omega_\nu}$
T_c
$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi} \sum_{\nu=0}^{\frac{\omega_b-\omega_a}{\Delta\omega}} \frac{Im\,F}{\omega_\nu} (j\omega_\nu) \cos \omega_\nu t \Delta\omega$
t = 0
$\frac{\pi}{\omega_a}$
[10]
[16b]

FIG. 34 RESPONSE TO A STEP FUNCTION


D = A - B
$\frac{\varphi(t)}{\sigma_{ST}} = \frac{1}{2} A(0) + \frac{1}{\pi} \sum_{\nu=0}^{\frac{\omega_b-\omega_a}{\Delta\omega}} A(\omega_\nu) \frac{\sin[\omega_\nu(-t) + \alpha(\omega_\nu)]}{\omega_\nu} \Delta\omega$
[10a]
B
$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi} \sum_{\nu=0}^{\frac{\omega_b-\omega_a}{\Delta\omega}} \frac{Re\,F}{\omega_\nu} (j\omega_\nu) \sin \omega_\nu t \Delta\omega$
[16a]
t = 0
$\frac{\pi}{\omega_a}$

It is therefore apparent that Equations [16*a*] and [16*b*] should offer reasonably good approximations of system response to step input functions if one half the period of the base frequency (expressed as $\pi \div \omega_a$) is greater than twice the apparent settling time of the system.

The exact expressions for output response, in terms of the real and imaginary components associated with the transfer function of the system, are given as

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi} \int_0^\infty \frac{\text{Re } F(j\omega)}{\omega} \sin \omega t d\omega \quad \ldots\ldots\ldots [14]$$

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi} \int_0^\infty \frac{Im\ F(j\omega)}{\omega} \cos \omega t d\omega \quad \ldots\ldots [15]$$

The author indicates that these equations are useful in that frequency-variant phase angles are removed from sine terms, permitting more simple methods of evaluating integral functions of these terms. Equations [16*a*] and [16*b*] are approximate solutions for Equations [14] and [15], resulting from graphical integration.

Another method of evaluating Equations [14] and [15], suggested in the paper, is to apply these expressions to straight-line approximations of

$$\frac{\text{Re } F(j\omega)}{\omega} \quad \text{or} \quad \frac{Im\ F(j\omega)}{\omega}$$

between finite-frequency limits. Where harmonic synthesizers are not available, this may be the preferred technique. For the same degree of accuracy, in many cases, fewer sine terms would be required in the approximation of system response, since "graphical" methods of integration require that the frequencies of the terms used in the approximation be equally distributed over the entire frequency range of the system. When a straight-line approximation is applied to $Im\ F(j\omega)/\omega$ over the entire frequency range of the system, the resulting solution for output response may be developed, simply, as follows

$$\int_0^\infty \frac{Im\ F(j\omega)}{\omega} \cos \omega t d\omega = Im\ F(j\omega) \frac{\sin \omega t}{\omega t}\Bigg]_0^\infty - \int_0^\infty \left(\frac{\sin \omega t}{t}\right) \frac{d}{d\omega} \left[\frac{Im\ F(j\omega)}{\omega}\right] d\omega$$

For the majority of cases where

$$Im\ F(j\omega) = 0 \text{ at } \omega = 0; \text{ and } \omega = \infty$$

and, noting that

$$\lim_{\omega t \to 0} \left[\frac{\sin \omega t}{\omega t}\right] = 1; \quad \lim_{\omega t \to \infty} \left[\frac{\sin \omega t}{\omega t}\right] = 0$$

$$Im\ F(j\omega) \frac{\sin \omega t}{\omega t}\Bigg]_0^\infty = 0$$

$$\int_0^\infty \frac{Im\ F(j\omega)}{\omega} \cos \omega t\, d\omega = -\int_0^\infty \frac{\sin \omega t}{t} \frac{d}{d\omega} \left[\frac{Im\ F(j\omega)}{\omega}\right] d\omega$$

Applying the foregoing to the curve of $Im\ F(j\omega)/\omega$ shown in Fig. 35, herewith, which is approximated by k straight lines

$$\int_0^\infty \frac{Im\ F(j\omega)}{\omega} \cos \omega t\, d\omega = -N_1 \int_0^{\omega_1} \frac{\sin \omega t}{t} d\omega - N_2 \int_{\omega_1}^{\omega_2} \frac{\sin \omega t}{t} d\omega \ldots - N_{k-1} \int_{\omega_{k-2}}^{\omega_{k-1}} \frac{\sin \omega t}{t} d\omega - N_k \int_{\omega_{k-1}}^{\infty} \frac{\sin \omega t}{t} d\omega$$

Thus

$$\int_0^\infty \frac{Im\ F(j\omega)}{\omega} \cos \omega t\, d\omega = \frac{1}{t^2} [-N_1 + (N_1 - N_2) \cos \omega_1 t + (N_2 - N_3) \cos \omega_2 t + \ldots (N_{k-1} - N_k) \cos \omega_{k-1} t + N_k \cos \omega_k t]$$

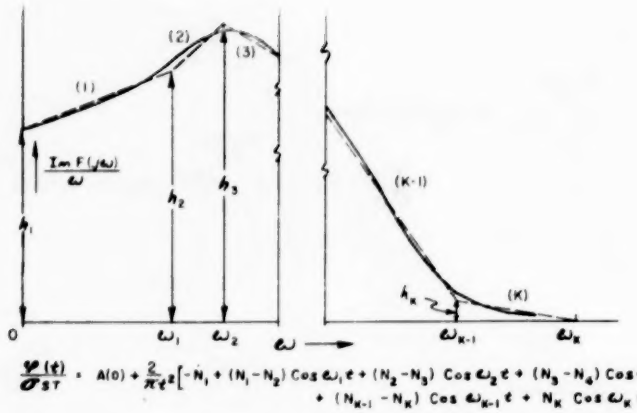


FIG. 35 STRAIGHT-LINE APPROXIMATION

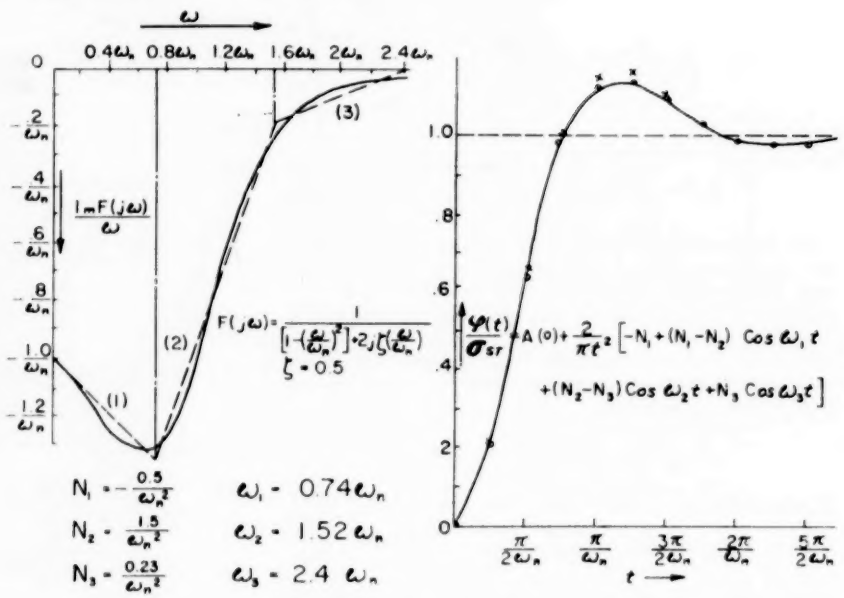


FIG. 36 SECOND-ORDER SYSTEM

where

N_1, N_2, etc. = Slope of line (1); line (2); etc.

$$= \frac{h_2 - h_1}{\omega_1}, \frac{h_3 - h_2}{\omega_2 - \omega_1}, \text{etc.}$$

Substituting into Equation [15][16]

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi t^2}[-N_1 + (N_1 - N_2)\cos\omega_1 t + (N_2 - N_3)\cos\omega_2 t + \ldots + (N_{k-1} - N_k)\cos\omega_{k-1} t + N_k \cos\omega_k t]$$

[16] Using

$$\lim_{\omega t \to 0} \cos\omega t = \left[1 - \frac{(\omega t)^2}{2}\right],$$

$$\lim_{t \to 0} \frac{2}{\pi t^2}[-N_1 + (N_1 - N_2)\cos\omega_1 t + (N_2 - N_3)\cos\omega_2 t + \ldots N_k \cos\omega_k t]$$

$$= \frac{1}{\pi}[(N_2 - N_1)\omega_1^2 + (N_3 - N_2)\omega_2^2 + \ldots - N_k\omega_k^2]$$

Fig. 36 of this discussion illustrates the accuracy of the method when applied to a second-order system with frequency-response characteristics illustrated in Fig. 32 of this discussion. Its curve of $Im\,F(j\omega)/\omega$ is shown at the left. In this case only three lines were used in the approximation. The predicted transient-recovery curve is shown at the right. Points on the curve of the exact solution are indicated by crosses. In this particular example, the two curves are shown to be almost identical.

AUTHOR'S CLOSURE

A. Tustin describes a harmonic analyzer by means of which it is possible to carry out the transformation from transient to frequency response, or the reverse transformation, as is necessary to solve the problems treated in the paper. Unfortunately, the author does not have such a device available, so that he is unable to evaluate the time that might possibly be saved by its use, as compared to other methods.

If the input to a system is made a unit impulse instead of a unit step, then we obtain the time functions of Equations [32] and [33] given in the discussion of C. H. Dawson. For the evaluation

of the transient, Floyd uses Equation [32], while Dawson makes use of Equation [33].

The author believes that the use of Equations [14] and [15] of his paper (which are valid for a unit-step input) is more practical for the approximate evaluation than the use of Equation [32] or [33]. The reasons are as follows:

1 If we calculate or construct both curves Re $F(j\omega)/\omega$ and $Im\ F(j\omega)/\omega$, then we can choose, for each given application, *that* curve which approaches zero most rapidly.

2 The curves corresponding to Equations [14] and [15] will always approach zero more rapidly than the curves corresponding to Equations [32] and [33].

3 The transient response to a unit step, which is what one is frequently interested in, is obtained directly by evaluating Equation [14] or [15]. When Equation [32] or [33] is used, an additional integration is still necessary before the desired result is obtained.

4 Equations [32] and [33] become unusable if the disturbance produces an immediate deviation of the controlled variable (as, for example, in Problem 1*b*, or a sudden inductive loading of a synchronous generator). The time derivative of the function $\varphi(t)$ then becomes infinite for $t = 0$, that is, $\varphi'(0) = \infty$.

If the graphical integration is broken off at the point $\omega_{\varphi i}$, $a_{\varphi i}$, then the resulting error can be estimated by means of Equation [20] or [21], depending on whether Re $F(j\omega)$ or $Im\ F(j\omega)$ is used for the evaluation. If Equation [21] is used, and if the integration is broken off at $\omega_{\varphi i}$ and $a_{\varphi i}$, then the largest error appears for $t = 0$ and is given by

$$\frac{\varphi_r(0)}{\sigma_{ST}} = \frac{2}{\pi}\, a_{\varphi i} \cdot \omega_{\varphi i}$$

since $\cos \omega t = 1$ for all values of ω when $t = 0$, and only when $t = 0$. If Equation [20] is used, and if the integration is broken off at $\omega_{\varphi r}$, $a_{\varphi r}$, then $\sin \omega t$ cannot be equal to 1 for all values of ω, no matter what value t may have. This, however, means that the error must be

$$\left|\frac{\varphi_r(t)}{\sigma_{ST}}\right| < \left|\frac{2}{\pi}\, a_{\varphi r} \cdot \omega_{\varphi r}\right|$$

Therefore, if the integration is broken off at ω_φ, a_φ, then the error will be equal to or smaller than $\frac{2}{\pi}\, a_\varphi \omega_\varphi$, in accordance with Equation [19]. The values $a_{\varphi i}$, $\omega_{\varphi i}$ and $a_{\varphi r}$, $\omega_{\varphi r}$ have no relation to each other, so that the Equation [34] of the discussion and the remarks associated with it lose their justification.

Direct determination of Re $F(j\omega)$ and $Im\ F(j\omega)$ of the closed loop from the open-loop data, as mentioned by Dawson, is only possible if we make definite assumptions about the point in the loop at which the disturbance (σ) acts. Therefore, the determination of Re $F(j\omega)$ and $Im\ F(j\omega)$ was carried out by means of a completely general method.

The method given in the discussion and represented by Equations [35] to [37] is, in the author's opinion, a valuable addition to the various possibilities for approximately solving Equations [14] and [15].

Dawson expands the curve Re $F(j\omega)/\omega$ into a Fourier sine series with the Fourier coefficients a_n, and then, using Equation [14], obtains the expression

$$\frac{\varphi(t)}{\sigma_{ST}} = 2\Omega \sin \Omega t \sum_{n=1}^{\infty} \frac{(-1)^{n+1} \cdot n \cdot a_n}{n^2\pi^2 - \Omega^2 t^2} \quad \ldots\ldots [37]$$

which can be evaluated relatively quickly by means of a table for the quantity

$$\left[(-1)^{n+1} \cdot \frac{2n \sin \Omega t}{n^2\pi^2 - \Omega^2 t^2}\right]$$

if the Fourier coefficients have been determined.

Since, in many cases, the curve $Im\ F(j\omega)/\omega$ is used for evaluating the response, and since this curve can usually be represented better by means of a cosine series, it might be worth while to show briefly how the method given by Dawson can also be adapted to such a case. Using Equation [15], we obtain the Fourier coefficients and the response given by

$$b_{i\omega} = \frac{1}{\Omega}\int_0^{\Omega} \frac{Im\ F(j\omega)}{\omega}\, d\omega$$

$$b_{in} = \frac{2}{\Omega}\int_0^{\Omega} \frac{Im\ F(j\omega)}{\omega} \cos\left(\frac{n\pi\omega}{\Omega}\right) d\omega$$

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \Omega b_{i\omega}\left[\frac{2 \sin (\Omega t)}{\pi(\Omega t)}\right] + \Omega \sum_{n=1}^{\infty}\left[(-1)^{n+1} \cdot \frac{2(\Omega t) \sin (\Omega t)}{\pi(n^2\pi^2 - \Omega^2 t^2)}\right] b_{in} \quad \ldots\ldots [40]$$

By means of tables for the quantities in brackets, this expression can also be evaluated relatively quickly.

S. R. Sporn uses a different way than Dawson (namely by means of a theorem by Shannon) to reach the same Equation [37], which, as has already been stated, has great advantages for the evaluation of the response.

In order to provide a better understanding of relations that have been obtained, E. C. Grogan points out how a periodic function can be changed into a unit step by means of a limiting process ($\omega_a \to 0$). If we do not use a unit step as the input to our system, but instead use a square wave with a fundamental frequency ω_a, and if we make sure that the time $\frac{\pi}{\omega_a} \gg T_e$ (where T_e is the apparent settling time of the system) then we also obtain a sufficiently accurate approximation of the transient response. Using this approach, Grogan succeeds in deriving Equations [14] and [15] in a relatively simple manner.

Treating the common case where $Im\ F(j\omega) = 0$ for $\omega = 0$ and for $\omega = \infty$, Grogan shows that, if the curve $Im\ F(j\omega)/\omega$ is approximated by straight-line segments, the transient response $\varphi(t)/\sigma_{ST}$ can be represented by means of simple cosine terms. Expanding this idea, and using Equation [17*a*] or [17*b*], the following relations for $\varphi(t)/\sigma_{ST}$ are obtained:

Using Re $F(j\omega)/\omega$ to evaluate the response, we have

$$\frac{\varphi(t)}{\sigma_{ST}} = \frac{2}{\pi} \cdot \frac{1}{t}\,[(a_1 - b_1) \cos \omega_1 t + (a_k + b_k) \cos \omega_{k+1} t] + \frac{2}{\pi} \cdot \frac{1}{t^2}\,[-N_1 \sin \omega_1 t + (N_1 - N_2) \sin \omega_2 t + \ldots + (N_{k-1} - N_k) \sin \omega_k t + N_k \sin \omega_{k+1} t] \quad \ldots\ldots [41]$$

The limiting value for $t = 0$ is given by

$$\left(\frac{\varphi(t)}{\sigma_{ST}}\right)_{t \to 0} = 0$$

Using $Im\ F(j\omega)/\omega$ to evaluate the response, we have

$$\frac{\varphi(t)}{\sigma_{ST}} = A(0) + \frac{2}{\pi} \cdot \frac{1}{t} \left[-(a_1 - b_1) \sin \omega_1 t + (a_k + b_k) \sin \omega_{k+1} t \right]$$

$$+ \frac{2}{\pi} \cdot \frac{1}{t^2} \left[-N_1 \cos \omega_1 t + (N_1 - N_2) \cos \omega_2 t + \ldots . \right.$$

$$\left. + (N_{k-1} - N_k) \cos \omega_k t + N_k \cos \omega_{k+1} t \right] \ldots \ldots [42]$$

The limiting value for $t = 0$ is given by

$$\left(\frac{\varphi(t)}{\sigma_{ST}} \right)_{t \to 0} = a_1(\omega_2 - \omega_1) + a_2(\omega_3 - \omega_2) + \ldots + a_k(\omega_{k+1} - \omega_k)$$

It is assumed that Re $F(j\omega)/\omega$ or $Im\ F(j\omega)/\omega$ can be set equal to zero for $\omega_1 > \omega > \omega_{k+1}$. The terms a_n and b_n are defined by Fig. 8, and

$$N_n = \frac{2b_n}{\omega_{n+1} - \omega_n}$$

The Equation [42] becomes identical to the equation given by Grogan if $\omega_1 = 0$ and $b_k = -a_k$.

Analysis of Regulating Systems With Particular Reference to Speed Control

By R. H. MACMILLAN,[1] CAMBRIDGE, ENGLAND

The design of speed-regulating systems is discussed in relation to that of other closed-loop control systems. Frequency-response methods are found to form a common basis suitable for the analysis of all such systems. The derivation of transfer functions for linear and quasi-linear elements, together with their application to the analysis of regulating systems, is described, and a simplified gas-turbine governor is studied in greater detail. A summary of the latest work likely to be of importance to control engineers is given, together with extensive references, but it is argued that they can make the best use of the work of others in the field only when a common approach is adopted universally. It is therefore concluded that an elementary study of these methods should be included in the training of every engineering graduate.

1 Introduction

THE first extensive application of automatic control was to a speed-regulating system. This was the governor designed by James Watt to control his steam engine. A century passed before Clerk Maxwell published the first mathematical study of governors and their stability. It was almost half a century later still that Nyquist and his associates devised the more powerful methods of frequency-response analysis for the treatment of such systems.

These methods were applied originally to feedback amplifiers and position-control systems, rather than to the design of regulators, which are defined in this paper according to British practice as control systems which maintain automatically at a fixed value the magnitude of some variable (usually engine or turbine speed) against relatively high-frequency disturbances. Meanwhile another group of engineers had been using automatic devices to control processes, and had arrived at some of the same design methods by a different route and disguised under a completely different terminology.

Until comparatively recently governors and process controllers were designed without any detailed analysis of their action when considered as part of a closed-loop system. This was possible because the control required of them was not to very close limits and because they did not need to be particularly quick-acting; that is, they did not have to respond to high frequencies. As the performance demanded becomes more exacting, however, and the sensitivity of the control is correspondingly increased, a stage is reached where self-oscillation of the system occurs. To obtain close control while avoiding the onset of such instability, a mathematical investigation becomes an essential preliminary to design, for only by this means is it possible to meet more stringent specifications. Such analysis cannot, of course, replace the development stage of a design, but it guides and supplements it. Frequency-response methods not only enable the designer to estimate the performance of a proposed system, which is the normal function of analysis; they also can assist him, to some extent, in the far more difficult problem of synthesis which aims to provide means for designing a system to meet a given specification. They even may go a step further and indicate the best performance obtainable within given physical and economic limitations.

The present situation is that analytical methods for linear and quasi-linear systems have outstripped practical development in all but the most advanced designs. The mathematical tools available to the designer are thus sufficient to enable him to make a full analysis of a linear control system provided that he can determine the necessary characteristics of its elements. The latter is one of the stumbling blocks which current research aims to remove. However, all engineers concerned with control systems will not be able to derive the full benefit from these powerful methods until it is generally appreciated that many of the problems faced by these different groups are fundamentally similar and that consequently, the same methods are suitable for their solution. Mathematical techniques provide just such a unifying basis, for they undoubtedly can lead not only to a uniform and rational approach to control problems, but also to a consistent nomenclature, based on the mathematical processes involved. This process of unification might be compared with that successfully achieved by the early chemists in bringing order out of the work of their alchemist predecessors, who used half a dozen different names for the same substance, according to the way it happened to have been made.

This approach, common to the study of all closed-loop systems, divides sharply into three stages. The first is the description of the physical system by means of a block diagram and transfer-function technique. The transfer function of an element describes concisely its frequency-response characteristics. If small disturbances only, about an equilibrium position of the system, are to be considered, it is possible to use a linearizing technique for some of the nonlinear elements. After analytical expressions have been derived for the transfer function of every element, an equation of motion for the system can be written down. It often happens in practice that the frequency response of an element is known only empirically. In such cases, either the best mathematical approximation may be taken as the transfer function or the available harmonic-response locus may be used directly without determining analytically an explicit equation of motion.

The next stage of this unified approach is the analysis of the information thus obtained. This enables one to predict the response of the system to various typical disturbances and thus to compare its performance with that of other systems, by applying suitable (or mathematically convenient) criteria. This completes the process of analysis. The final stage is synthetic; methods are available for altering the coefficients and form of the equation of motion in such a way as to improve the performance of the system. This process of adjustment will present con-

[1] University Lecturer, Department of Engineering, The Engineering Laboratory, Cambridge University.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29-December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 14, 1953. Paper No. 53—A-15.

siderably greater difficulties in its application to some systems than to others, but it is inherent in all rational design.

The aim of this paper is to amplify and justify some of the foregoing statements, the arrangement being as follows: In the next section is considered the relationship between the problems of speed regulation and those of process and position control; it is apparent that their similarities are more remarkable and important than their differences. Section 3 is a discussion of some of the most recent advances in the techniques of control-system analysis and synthesis. In section 4 the role of education in this field is examined, particular attention being paid to pregraduate teaching, which the author is best qualified to discuss. Section 5 includes some points of interest in connection with the use of transfer functions, particularly for quasi-linear systems; their effectiveness is exemplified, in section 6, by the systematic preliminary analysis of a simplified gas-turbine speed regulator. Linearized transfer functions are derived for its elements and displayed on a block diagram; the equation of motion is found and the response investigated; possible methods of improvement are then suggested. The last section is a brief discussion of some of the requirements associated with the regulation of hydraulic and gas turbines, together with the methods being used to meet them. The mathematical derivation of certain transfer functions is relegated to the Appendix.

2 Relation of Regulation Problems to Those in Other Control Fields

All control systems are governed by equations of motion which are liable to become inadequately damped or even unstable for certain values of the system parameters. Therefore there is no fundamental objection to the use of a uniform approach, both for obtaining the equation of motion and for its analysis.

In process control the main problems consist in finding an adequate representation for the behavior of the load and the practical difficulty of making a measurement of the value of the controlled variable, which will permit a direct comparison with its desired value. Various delays are usually involved in the latter process, which must vitiate the performance of the system as a whole. The time constants of the load and of the disturbances are, however, relatively large when compared with those of the controlling mechanisms, so that inertia in the controlling apparatus is of secondary importance and the speed of response need not be pushed to the limit.

In position control, on the other hand, the system must respond to a rapidly fluctuating input signal and, the sensitivity of the control being correspondingly high, the time lags in the servomechanisms are often the dominating factor. In addition, the control must attempt to distinguish between genuine input signals, requiring motion of the load, and spurious signals, usually of high frequency, superimposed upon them.

Speed regulation involves some of the problems of each of these fields. The response must be rapid, so that the sensitivity must be high and the considerable difficulties involved in the rapid measurement of speed, together with the delays inevitably associated with the changing of the engine power output, are apt to lead to instability. Until quite recently the problem of speed measurement has occupied too important a place in the minds of those concerned with engine governing. This is even more true of the textbooks at present available, which are concerned exclusively with the properties of governors and almost disregard the fact that the governor is only the measuring element in a complete closed-loop system. The emphasis, happily, is now shifting to the wider outlook.

The design of automatic pilots for aircraft affords an interesting example of a control with the dominant characteristics of a regulation problem, which has been handled successfully by the techniques normally reserved for position control. Nothing but good has come from this fusion of ideas.

These minor distinctions between the problems confronting control engineers in the various fields can no longer justify the radically different terminology which they still normally adopt. Progress in the past might have been more rapid had those concerned appreciated, for example, that an "inverse Nyquist diagram" and a "potential correction curve" are essentially the same thing (1).[2,3] It is time for general replacement of such primitive expressions as "temporary return motion," "transient reset" and "proportional-band width." Only when the ground has been cleared in this way will engineers be in a position to make full use in their own particular field of the various techniques for analysis and synthesis which have been devised during the past decade. Standard glossaries of terms, collating the different usages, already have been prepared both in Great Britain and in the United States.

3 Recent Advances in Analytical Techniques

The past few years have witnessed a number of striking advances in the technique of control-system analysis, the present tendency being to develop methods which can be extended to assist in synthesis.

Alternative forms of the harmonic-response locus, or Nyquist diagram, such as the inverse-response locus and the logarithmic or asymptotic methods of plotting, assist in studying the effects of feedback and variation of the system parameters. These have been developed further to give a graphical estimate of the roots of the characteristic equation, which are needed to determine the transient response (2). Sufficient information about the transient performance for most practical purposes can be obtained readily by computing certain of its salient characteristics on the assumption that there is a predominant mode of oscillation in the response of complex systems (3). This simplification is widely applicable, but several alternative methods of determining the transient response also have been devised (4, 5).

A method has been proposed for adjusting the coefficients of the equation of motion to give an adequately damped system (6). This extends the criterion of Routh, which can do no more than insure the presence of positive damping. Other methods for determining the limiting values of the parameters, such as the sensitivity and the time lags, which will give rise to continuous oscillation, have been extended to indicate any desired degree of damping (7, 8). The "root-locus" method, which involves a plot on the damping/frequency plane with gain as the variable parameter (9) has been generalized, by means of the "phase-angle locus" (10, 4) and its usefulness thereby greatly increased.

Stability charts for a rather general system subject to a randomly varying input have been produced (11) and, stemming from Wiener's work, the problem of optimum response, on the basis of minimizing the average of error squared (12) has led to the new concept of "form optimization," the best functional form for the transfer functions being determined. The analytical and practical possibilities of various other criteria for optimum response also have been investigated. These include the minimization of the integral of the modulus of the error (13) in a transient and a time-weighted function of the error (14, 15, 16). Another approach to transient-response optimization permits the inclusion of other minimal criteria and leads, in general, to nonlinear systems (17).

Application of frequency-response methods to the analysis of

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

[3] It is also closely related to the property often described as "mechanical impedance."

nonlinear systems has led to one fundamental development, whereby the concept of a transfer function is generalized to that of a "describing function" (19), which is plotted in the gain/phase plane with signal amplitude as the variable parameter. Meanwhile, the "phase-plane" method, with displacement plotted against velocity and time as the variable parameter, has been applied successfully to the analysis of discontinuous systems (18, 20). A graphical construction for the phase-plane locus permits the rapid determination of the transient response to an arbitrary input function of nonlinear second-order equations of almost unlimited complexity (21). Discontinuous systems of the relay type also have been studied with a view to obtaining optimum characteristics, and much attention has been paid to step-by-step and sampling systems. However, these are beyond the scope of the present survey. A simulator study, in which the high-order transfer-function characteristic of many processes is approximated by means of a finite delay together with a simple time lag (22), suggests considerable possibilities for future analytical simplification in this field.

4 The Role of Education

The techniques mentioned in the foregoing section obviously are suitable only for postgraduate study, where courses on control-system analysis are becoming of ever-increasing importance in meeting their obligation to prevent the appearance of a gap between theoretical investigations and the practical application of their results.

It is in the elementary courses, leading to a first engineering degree, that sufficient attention is too often not paid to teaching the basic principles of closed-loop system operation, those principles which are common to all feedback controls. Of the three stages of the common approach detailed in section 1, the whole of the first stage and some part of the second should be included. The necessary mathematical knowledge is not considerable and certainly should be at the command of every graduate engineer.

It is already normal practice to include instruction in control in the courses for all electrical and chemical engineers, but such instruction is not yet universally adopted for mechanical engineers. In Cambridge, England, we have found it convenient to devote a part of the "Theory of Machines" course to this topic, arising naturally as it does out of the problems of speed governing.

An introduction of this nature is not only of intrinsic value, for every engineer must frequently come into close contact with automatic controls, but in the practical applications it should act as a unifying influence between various fields which tend, with the present trend toward excessive specialization, to become unduly separated.

5 Transfer Functions and Block Diagrams

The transfer function of a linear or quasi-linear element is the differential relation connecting the output signal from it with the input signal to it. Unless a separate feedback loop is incorporated in the block diagram, it is important that the input should be unaffected by the output, for the transfer function should represent a passage of information across the element in one direction only.

The transfer function must have one of six standard forms or be a combination of one or more of them (2). It is intimately related to the harmonic response, for if $p = d/dt$ is put equal to $j\omega$, the resulting expression gives the amplitude and phase change of the response to a unit sine wave of frequency $\omega/2\pi$. The harmonic response is thus immediately available if the form of the transfer function can be deduced from physical considerations.

The reverse process of assigning a transfer function to an element with a known harmonic-response locus is more difficult, but it can be performed by means of various approximations. Often, however, no such mathematical representation is necessary for the analysis of a particular system to proceed. By taking account only of the first harmonic component of the response, the transfer loci for a nonlinear element (the describing loci) can be plotted similarly, but a separate locus is required for every signal amplitude. Fortunately, in important practical applications, the loci for different frequencies overlap, in which case a single describing locus remains, with signal amplitude as the variable parameter.

In the early stages of design it is convenient to find mathematical expressions for the transfer functions from theoretical considerations, by writing down from first principles the equation connecting the input X and output Y of the element. Suppose this relation to be of the nonlinear form

$$f_1(Y) = f_2(X, N) \quad [1]$$

where N is another variable parameter. Then, differentiating, we find

$$\frac{\partial f_1}{\partial Y}\,\Delta Y = \frac{\partial f_2}{\partial X}\,\Delta X + \frac{\partial f_2}{\partial N}\,\Delta N \quad [2]$$

as the linear relationship connecting small finite increments ΔY, ΔX, and ΔN of the variables. If, for example, we had

$$Y^2 + A = BX\sqrt{N} \quad [3]$$

where A and B are constants, then

$$2Y\,\Delta Y = (B\sqrt{N})\Delta X + \left(\frac{1}{2}BX/\sqrt{N}\right)\Delta N \quad [4]$$

A convenient notation is to use lower-case symbols for small finite increments, writing y for ΔY, x for ΔX, and n for ΔN. The values of X, Y, and N in Equation [4] are the mean values about which the disturbance is taking place. Calling these X_0, Y_0, and N_0, Equation [4] becomes

$$2Y_0\,y = (B\sqrt{N_0})\,x + \left(\frac{1}{2}BX_0/\sqrt{N_0}\right)n \quad [5]$$

which is the linearized relationship between x, y, and n, applicable only for small values of these variables.

In systems with more than one degree of freedom it may be convenient to use a generalized transfer function or "transfer matrix." Consider, for example, the speed and temperature control of a turbopropeller aircraft engine. The linearized relation for the dependence of engine-speed increment n, on propeller-blade-angle increment β, and fuel-flow increment q, is

$$n + Tpn = k_1q - k_2\beta \quad [6]$$

where T is the rotor time constant. The temperature change θ also depends on these variables as follows

$$\theta = k_3q - k_4n \quad [7]$$

From these equations we obtain the relations

$$n = \frac{k_1}{1 + Tp}\,q - \frac{k_2}{1 + Tp}\,\beta = aq + b\beta \quad [8]$$

and

$$\theta = \frac{k_3(1 - k_1k_4/k_3 + Tp)}{1 + Tp}\,q + \frac{k_2k_4}{1 + Tp}\,\beta = cq + d\beta \quad [9]$$

The engine transfer matrix, relating n and θ the dependent variables, to q and β the independent variables, is therefore

$$\begin{matrix} a & b \\ c & d \end{matrix}$$

Such transfer matrixes can be manipulated according to the usual rules of matrix algebra and their use is likely to facilitate the design of complex systems with several degrees of freedom.

When the system being treated has only one degree of freedom, it can be represented by a block diagram containing only simple transfer functions and adding elements. The equation of motion is then written down from the diagram. It may be useful at this stage to nondimensionalize the variables by expressing them as proportions of their steady values. Thus dividing Equation [5] by $X_0Y_0N_0$ throughout and writing $\hat{x} = x/X_0$, and so on, for the nondimensional variables, we obtain on rearrangement

$$2[1 - A/(Y_0^2 + A)]\,\hat{y} = \hat{x} + \frac{1}{2}n \quad \ldots\ldots [10]$$

using the fact that

$$Y_0^2 + A = BX_0\sqrt{N_0} \quad \ldots\ldots [11]$$

Although the dimensionless technique is invaluable for generalizing results in the later stages of an analysis, it is not normally to be recommended initially, as its use is apt to make it difficult or impossible to obtain dimensional checks on the accuracy of equations.

6 Gas-Turbine Speed Governor

There follows a preliminary analysis, involving the necessary approximations stated, of a speed regulator for a gas turbine. The system is nonlinear but can be linearized to investigate the nature of the response for small disturbances about a stable operating condition. It is a single-degree-of-freedom system.

Control is effected by varying the quantity of fuel pumped to the burners. The speed is indicated by a measurement of the pressure increase in the impeller, the low pressure from the center and the centrifugal pressure at the periphery being fed to opposite sides of a diaphragm whose displacement opens the orifice of a relay operating the fuel by-pass valve. As the speed increases, the centrifugal pressure rises and the increased opening of the orifice allows a larger proportion of the fuel being pumped to be bypassed, so that the supply to the burners is diminished and the speed thus tends to fall again.

A descriptive block diagram for the system is shown in Fig. 1. Each block is characterized by a linear transfer function, which is a function of $p = d/dt$. These functions are denoted by

$f_m(p)$ for measurement
$f_r(p)$ for relay
$f_v(p)$ for valve and by-pass
$f_b(p)$ for burners
$f_l(p)$ for rotor and load

Now the system has been reduced to the standard form, it is next necessary to find analytical expressions for the various transfer functions involved.

Impeller and Diaphragm. The arrangement is shown diagrammatically in Fig. 2. Ignore any time lag caused by the pipe lines

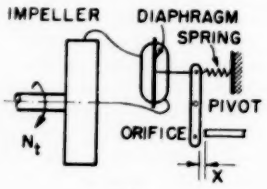


Fig. 2

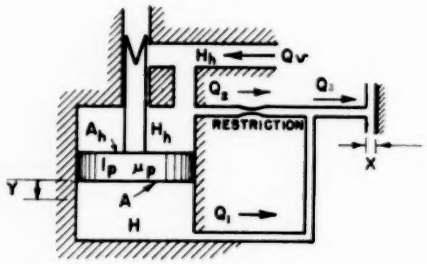


Fig. 3

from the impeller to the diaphragm and assume that the pressure difference in the impeller changes instantly with change in turbine speed, N_t. N_d is the desired (set) speed.

With these assumptions it can be shown that, using the notation of section 5 for small variations

$$x = K_m(n_t - n_d) \quad \ldots\ldots [12]$$

The derivation, together with the relationship of the constant to the physical properties of the elements, is given in the Appendix. It has thus been proved that

$$f_m(p) = K_m \quad \ldots\ldots [13]$$

Pilot Relay. This is shown in Fig. 3. As the orifice aperture increases with X, the pressure H drops, and the piston is forced downward, thus allowing an increased flow of fuel to be bypassed. Assuming that the flow through the orifice never becomes very small, it is shown in the Appendix that

$$(1 + T_r p)py = K_r x \quad \ldots\ldots [14]$$

whence

$$f_r(p) = \frac{K_r}{p(1 + T_r p)} \quad \ldots\ldots [15]$$

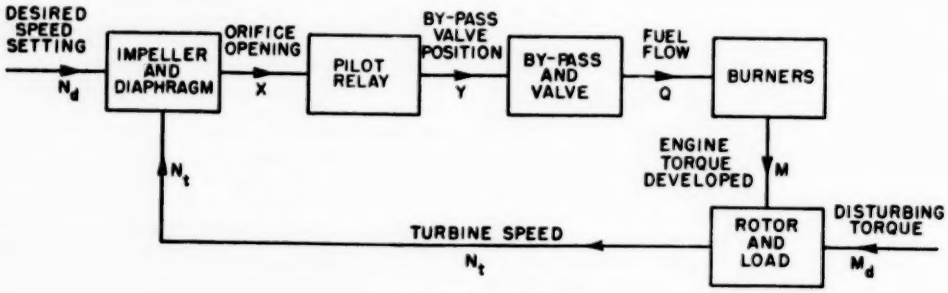


Fig. 1

the time constant T_r being directly proportional to the mass of the relay piston.

Valve and By-Pass. Let Q_p be the constant flow of fuel discharged by the pump and Q_v the amount by-passed, as shown in Fig. 4. Then the flow to the burners is

$$Q = Q_p - Q_v \quad [16]$$

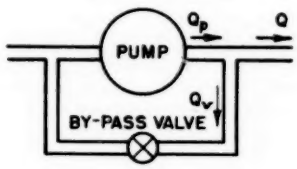


FIG. 4

If the profile of the by-pass valve is such that

$$Q_v = \phi(Y) \quad [17]$$

the pressure difference across it being practically constant, then

$$Q = Q_p - \phi(Y) \quad [18]$$

or, for small variations

$$q = -(d\phi/dY)y = -K_v y \quad [19]$$

so

$$f_v(p) = -K_v \quad [20]$$

Burners. If there is a delay time T_b in the combustion chamber between the increase in fuel flow and the development of an increased torque M, then

$$(1 + T_b p)m = K_b q \quad [21]$$

so

$$f_b(p) = \frac{K_b}{1 + T_b p} \quad [22]$$

Rotor and Load. The equation of motion for the turbine expresses the fact that the torque available for accelerating the rotor, of polar inertia I, is equal to the internally developed torque less the sum of the disturbing torque and the resistive torque, taken as μ times the angular velocity, that is

$$IpN_t = M - M_d - \mu N_t \quad [23]$$

whence

$$n_t = \frac{m - m_d}{\mu + Ip} \quad [24]$$

so

$$f_l(p) = \frac{K_l}{1 + T_l p} \quad [25]$$

where

$$K_l = 1/\mu \text{ and } T_l = I/\mu \quad [26]$$

If the viscous friction is relatively small or its effective value is reduced by the load characteristics, e.g., constant power consumption, then μ can be neglected and

$$f_l'(p) = K_l'/p \quad [27]$$

with

$$K_l' = 1/I$$

Equation of Motion. It is now possible to set down the analytical block diagram shown in Fig. 5. The governing equation for the system is immediately written from it, as

$$n_t = f_l(m - m_d) = f_l f_b f_v f_r f_m (n_t - n_d) - f_l m_d \quad [28]$$

or

$$(f_l f_b f_v f_r f_m - 1) n_t = f_l f_b f_v f_r f_m n_d + f_l m_d \quad [29]$$

Substituting in Equation [29] the expressions found for the transfer functions, we have for the governing equation

$$[p(1 + T_r p)(1 + T_l p)(1 + T_b p) + K] n_t = K n_d - K_l p(1 + T_r p)(1 + T_l p) m_d \quad [30]$$

where $K = K_l K_b K_v K_r K_m$, the loop gain. Clearly, in the steady state, where $p = 0$, $n_t = n_d$, independently of the value of the disturbing torque m_d.

If the output damping μ is small, the load gain K_l and time constant T_l become large and the system will tend to become unstable. Replacing f_l by f_l', the system is seen to have two integrations in the loop, which leads to the same conclusion. Stability could be restored either by turning the integrating relay into a proportional one or by adding a derivative term ("phase advance") to the signal in the loop. The first method could be achieved either by means of a spring opposing the piston motion or by introducing a local feedback in the relay. A quadratic transfer function results in either case, as shown in the Appendix. This method of stabilizing, though very convenient, would permit a drop in steady speed with increase of load torque, when the small though finite load damping is taken into account. The other method would involve some modification of the relay or the addition of a phase advance element, but would not introduce a steady-state error.

In order to choose the system parameters, a more detailed analysis, by any of the methods enumerated in section 3, can be applied to the governing equation, but it must be borne in mind that the results are valid only for small variations about the operating speed. For a sudden large increase in load torque, the orifice of the relay will be closed completely and the system will then operate as an open loop until the speed has again increased sufficiently for the orifice to open again.

The same argument applies to the response when the set speed is changed suddenly from idling to the operating speed. The by-pass valve would be closed completely by the time the speed

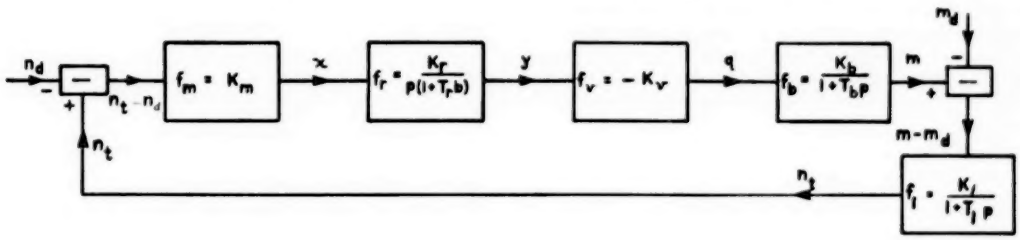


FIG. 5

has risen enough to open the orifice and, unless there is considerable phase advance, an overshoot would be expected to occur before the relay again opens the by-pass and cuts down the fuel supply. The linearizing assumptions made previously for the orifice would not be valid under these circumstances and a step-by-step computation would be necessary to estimate the amount of the overshoot. It is therefore not possible to generalize about the behavior under such circumstances.

7 Turbine Regulation

It is perhaps worth mentioning, finally, some of those factors particularly associated with problems of turbine regulation which may complicate their mathematical formulation and subsequent analysis. Most of the work on hydraulic turbines has been done in countries where hydroelectric power is a major industry. Further references will be found in the two papers cited (**23**, **24**).

One problem arises from the means used to effect control of the power output of the turbine. With hydraulic turbines it is generally necessary to use some form of double regulation. With a Pelton wheel, for example, the main control is by means of a spear in the nozzle, which reduces the flow of water as it moves forward. This movement must not be too rapid, or too great a pressure rise would occur in the duct leading to the nozzle. On a sudden reduction of load an unacceptable increase in speed would therefore occur if the spear were the only means of controlling the flow, and it is thus necessary to provide some sort of temporary flow diversion by means of either a quick-acting diffuser or deflector.

Another application involving a multiloop system is when the governor is required to control some other variable beside speed. One example is the hydraulic-turbine governor, which controls the level of head or tail water; another example is the gas turbine, mentioned in section 5, with temperature limitation and control of both fuel flow and propeller-blade angle. In many such cases it is possible to regard one control as constituting the main loop, variations in the other being temporarily ignored in its analysis, the disturbing effect of the subsidiary loop being considered subsequently.

Even without temperature control a satisfactory aircraft-engine turbine requires a more elaborate method of control than that discussed in the last section. The pilot's throttle lever must have a more direct control over the fuel flow than adjustment of the spring force in the pressure diaphragm would give. In practice it varies an orifice in the line to the burners, while a subsidiary control, acting in conjunction with the by-pass relay, maintains the constancy of the pressure drop across the throttling orifice. The constant-speed regulation is then only effective when the speed is in the region of full throttle speed, one of its most important functions being to act as an overspeed governor, preventing excessive speed in rapid acceleration from idling.

In deriving the transfer function for a reaction-turbine control system, it is necessary to take into account the effect of water hammer in increasing the hydraulic supply pressure as the guide vanes close. If T_c, the time constant of the hydraulic conduit, is defined as the ratio of the kinetic energy of the water in the conduit to the power at the nozzle and T_r, the time constant of the runner, is defined as the ratio of the kinetic energy of the runner to the power output from it, then the transfer function relating change in runner acceleration to the guide vane movement γ, has the form (**24**)

$$\frac{pn}{\gamma} = \frac{1 - T_c p}{T_r\left(1 + \frac{1}{2}T_c p\right)}$$

from which it is apparent that the water hammer has an unstabilizing influence.

Conclusions

The problems confronting the designers of speed-regulating controls are for the most part similar to those faced by other control engineers, and the mathematical methods of analysis now available are applicable to all control systems. It is the function of engineering education, particularly in the elementary stages, to insure that all engineers adopt a uniform approach to these problems. Mechanical, chemical, and electrical engineers should speak a common technical language.

The advances in mathematical technique which have taken place in the past few years are considerable, and one of the prime requirements at present is that engineers shall be able to express their problems in a suitable form to make full use of these new techniques. The difficulty, of course, is by no means limited to that of education; it is in many instances inherent in the problems themselves. The important thing in such cases is to investigate those factors which are preventing the particular problem from being related to the general fund of control knowledge. It is in this direction that the author foresees the immediate progress of practical control-system design.

BIBLIOGRAPHY

1 "The Frequency Response Method: Its Relationship to Transient Behaviour in Control System Design," by J. H. Westcott, Transactions of the Society of Instrument Technology, vol. 4, 1952, p. 113.

2 "An Introduction to the Theory of Control in Mechanical Engineering," by R. H. Macmillan, Cambridge University Press, Cambridge, England, 1951.

3 "Synthesis of Feedback Control Systems by Phase-Angle Loci," by Y. Chu, Trans. AIEE, part II, vol. 71, 1952, p. 330.

4 "Comparison of Steady State and Transient Performance of Servomechanisms," by H. Chestnut and R. W. Mayer, Trans. AIEE, vol. 68, part I, 1949, p. 765.

5 "Transient Performance From Frequency Response Plots," by H. Harris, M. J. Kirby, and E. F. von Arx, Trans. AIEE, vol. 70, part II, 1951, p. 1452.

6 "Theory of Servomechanisms With Particular Reference to Stabilization," by A. L. Whiteley, *Journal of the Institution of Electrical Engineers*, vol. 93, 1946, p. 353.

7 "Relative Damping as Criterion for Stability," by A. Leonhard, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, A. Tustin, editor, 1952, p. 25.

8 "A Generalization of the Frequency Response Method for the Study of Feedback Control Systems," by N. L. Kusters and W. J. M. Moore, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, A. Tustin, editor, 1952, p. 105.

9 "Control System Synthesis by Roots Locus Method," by W. R. Evans, Trans. AIEE, vol. 69, 1950, p. 66.

10 "Feedback Control Systems With Dead Time or Distributed Lag," by Y. Chu, Trans. AIEE, vol. 71, 1952, p. 291.

11 "The Human Transfer Function in Servo-Systems," by J. D. North, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, p. 473: see explanatory discussion by A. Tustin, p. 498.

12 "The Development of Relationships Concerning the Frequency Bandwidth and Mean Square Error of Servomechanisms From Properties of Gain-Frequency Characteristics," by J. H. Westcott, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, p. 45.

13 "The Calculation of the Optimum Parameters for a Following System," by C. Mack, *Philosophical Magazine*, London, England, vol. 40, 1949, p. 922.

14 "Some Design Criteria for Automatic Controls," by P. T. Nims, Trans. AIEE, vol. 70, 1951, p. 606.

15 "Controller Settings for Optimum Control," by W. A. Wolfe, Trans. ASME, vol. 73, 1951, p. 413.

16 "Theoretical Considerations on the Optimum Adjustment of Regulators," by P. Hazebroek and B. L. van der Waerden, Trans. ASME, vol. 72, 1950, p. 309.

17 "Automatic Control Systems Satisfying Certain General Conditions on Transient Behaviour," by A. S. Boksenbom and R. Hood, NACA Report 1068, 1952.

18 "The Stabilisation of On-Off Controlled Servomechanisms," by A. M. Uttley and P. H. Hammond, *Automatic and Manual Control*,

Butterworths Scientific Publications, London, England, 1952, p. 285; see also discussion, p. 300.

19 "Sinusoidal Analysis of Feedback Control Systems Containing Non-Linear Elements," by C. E. Johnson, Trans. AIEE, vol. 71, 1952, p. 160.

20 "The Analysis of Relay Servomechanisms," by H. K. Weiss, *Journal of the Aeronautical Sciences*, vol. 13, 1946, p. 364.

21 "On the Graphical Solution of Transient Vibration Problems," by R. E. D. Bishop, to appear in the Proceedings of The Institution of Mechanical Engineers.

22 "On the Automatic Control of Generalized Passive Systems," by Kun Li Chien, J. A. Hrones, and J. B. Reswick, Trans. ASME, vol. 74, 1952, pp. 175–185.

23 "Water Turbine Governors," by N. G. Dennis, Proceedings (B) of The Institution of Mechanical Engineers, vol. 1B, 1953, pp. 379–400.

24 "Electro-Hydraulic Control of Water Turbines," by A. Garde, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, p. 503.

Appendix

Mathematical Derivation of Transfer Functions

Impeller and Diaphragm. Referring to Fig. 2, let P_d be the pressure difference across the diaphragm; then with the notation and assumptions given in section 6, Impeller and Diaphragm

$$P_d = k_1 N_t^2 \quad [31]$$

Differentiating this and putting $\Delta P_d = p_d$, and so on

$$p_d = 2k_1 N_t n_t \quad [32]$$

For equilibrium of the diaphragm

$$k_2 p_d = k_3 X + F_1 = k_3 X + \lambda_1 N_d \quad [33]$$

where F_1 is $\lambda_1 N_d$, the set force in the spring. Differentiating again

$$k_2 p_d = k_3 x + \lambda_1 n_d \quad [34]$$

Combining with Equation [32]

$$x = \frac{2k_1 k_2 N_t}{k_3} n_t - \frac{\lambda_1}{k_3} n_d \quad [35]$$

To obtain good governing, $n_t = n_d$ when $x = 0$; therefore

$$\lambda_1 = 2k_1 k_2 N_t \quad [36]$$

and

$$x = K_m(n_t - n_d) \quad [37]$$

where

$$K_m = \frac{2k_1 k_2 N_t}{k_3} \quad [38]$$

Integrating Relay. According to the notation in Fig. 3, the pressure difference across the piston is $H_b - H$; its inertia is I_p and its motion is opposed by viscous friction μ_p times its velocity. Then, for the flows

$$Q_3 = Q_2 + Q_1 \quad [39]$$

where

$$Q_1 = A \frac{dY}{dt} \quad [40]$$

and

$$Q_2 = c_1 \sqrt{(H_b - H)} \quad [41]$$

Differentiating Equations [39] and [40] and eliminating q_1 we obtain

$$q_3 = q_2 + Apy \quad [42]$$

Differentiating Equation [41], we have

$$q_2 = -k_4 h \quad [43]$$

where

$$k_4 = c_1/2\sqrt{(H_b - H)} \quad [44]$$

For the orifice, we have

$$Q_3 = c_2 X \sqrt{H} \quad [45]$$

Differentiating

$$\Delta Q_3 = c_2 \sqrt{H} \Delta X + \frac{c_2 X}{2\sqrt{H}} \Delta H$$

or

$$q_3 = k_5 x + k_6 h \quad [46]$$

where

$$k_5 = c_2 \sqrt{H} \text{ and } k_6 = \frac{1}{2} C_2 X / \sqrt{H} \quad [47]$$

For the motion of the piston the following relation must always hold

$$I_p p^2 Y + \mu_p p Y = A_b H_b - AH \quad [48]$$

whence

$$(I_p p^2 + \mu_p p) y = -Ah \quad [49]$$

eliminating q_2, q_3, and h from Equations [42], [43], [46], and [49]

$$\frac{Ak_5}{k_4 + k_6} x = \left[I_p p + \left(\mu_p + \frac{A^2}{k_4 + k_6} \right) \right] py \quad [50]$$

so

$$py = \frac{K_r x}{1 + T_r p} \quad [51]$$

where

$$K_r = \frac{Ak_5}{k_4 + k_6} \Big/ \left(\mu_p + \frac{A^2}{k_4 + k_6} \right)$$

and

$$T_r = I_p \Big/ \left(\mu_p + \frac{A^2}{k_4 + k_6} \right)$$

Proportional Relay. If a spring of stiffness k_s controls the relay piston, Equation [49] becomes

$$(I_p p^2 + \mu_p p + k_s) y = -Ah \quad [52]$$

and Equation [50] is then

$$\frac{Ak_5}{k_4 + k_6} x = \left[I_p p^2 + \left(\mu_p + \frac{A^2}{k_4 + k_6} \right) p + k_s \right] y \quad [53]$$

The relation between y and x is then a proportional one, with a quadratic delay.

The same effect is obtained if the orifice opening, x_1, is controlled through a differential lever, both from the diaphragm movement x, and from a proportion of the piston movement, by means of a linkage. The equations for the relay are then

$$py = \frac{K_r x_1}{1 + T_r p} \quad [54]$$

and

$$ay + bx_1 = cx \quad [55]$$

Elimination of x_1 between these gives an equation of the same form as Equation [53].

Discussion

R. Oldenburger.[4] In the matter of education the role of mathematics is extremely important. The same type of mathematics applies to different areas of application and as the author so aptly points out, it is the mathematics that is the unifying agent. Intelligent use of modern control theory requires an extensive knowledge of the vast body of underlying mathematics. An understanding of Dr. Nyquist's pioneering frequency-response paper alone requires a considerable background in the theory of functions of complex variables. The more mathematics a control expert knows the better able he should be to find quick and efficient methods of solving his problems and to employ analytical techniques that fit the specific problem.

These facts suggest that it is desirable for the future automatic-control expert to major in mathematics both in his undergraduate and graduate work and to minor in the physical sciences and engineering. It is unquestionably much easier for a person with an extensive mathematical background to pick up the engineering, such as the physical limitations involved in the design of equipment, than the other way. It is very important that the control expert be trained in engineering in general, and not be limited to electrical, mechanical, or other specific branches of the field. Training in the use of computers is also essential.

The various transfer functions given by the author are in common use in the design of speed governors for prime movers. They are naturally valid for rough studies and noncritical areas, but will break down in certain situations. Thus we found that the gas-turbine equations of the paper were inadequate for explaining the performance of our governors on these turbines; in fact we discovered experimentally by frequency-response runs (verified later by NACA) that we could not neglect a certain combustion lag that appeared as a dead time. This lag could be as much as $^1/_4$ or $^1/_2$ sec. In this case the denominator in Equation [22] should include a factor e^{Tp} for the dead time T. This dead time T is in addition to the delay time T_b of Equation [22] and is due to the time it takes to go from one combustion pattern to another.

The gas turbine-governor application probably was given only for purposes of illustration. In practice we would measure the speed of the gas turbine directly. In fact, it is our experience that variables to be controlled should be measured as directly and accurately as possible. The arrangement of Fig. 2 of the paper shows a measurement of pressure. It is evident that there will not be a one-to-one correspondence between pressure and speed holding at each instant during a transient.

[4] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

Author's Closure

Dr. Oldenburger is undoubtedly correct in asserting that an extensive mathematical background will be of great assistance to any control expert. Surely this applies to most fields of modern engineering. The study of mathematics must form the basis of all higher technology. Nevertheless, the author believes it is important to distinguish between the control experts, the designers or researchers, and the far larger number of engineers who use or buy automatic instrumentation and machinery. It is valuable for them also to have a working knowledge of the theory of feedback systems, for which no very extensive mathematical equipment is required beyond an acquaintance with the properties of complex numbers and the ability to solve linear differential equations with constant coefficients. The author has attempted to show in his introductory article to the symposium publication that this much is all that is indispensable.

Undoubtedly a finite delay or dead time will occur in the loop of many control systems; fortunately it is not necessary to resort to approximate analysis when this happens, for exact methods are available to take account of this type of linear delay, using the transfer function given by Dr. Oldenburger. In reference (22) of the paper a method is given for approximating the transfer function of many processes by means of a combination of finite and exponential delays. Charts computed on this basis, to assist in the setting of process controllers, also might be used by the designers of governors, provided that suitable nondimensional scales have been used.

It is always important to measure the controlled variable of any servosystem as accurately as possible, and it is best to compare the actual value of the controlled variable as directly as possible with the desired value. Whether the output of the measuring device is to be a mechanical displacement or a change of electrical potential, some delay inevitably is involved in obtaining it, if its physical nature differs from that of the quantity measured. An exact one-to-one correspondence between turbine speed and the quantity representing the measurement is therefore never possible. It is no more direct to use a mechanical centrifugal governor than to use a measurement of centrifugal pressure, although in practice the time delay involved in the former certainly might be made much the shorter.

Application of Frequency-Analysis Techniques to Hydraulic Control Systems

By A. C. HALL,[1] DETROIT, MICH.

Hydraulic control systems, while inherently nonlinear, under many conditions can be studied by means of a linear approximation. Performance may be predicted by this approximation with reasonable accuracy provided the change in fluid flow during a transient is small compared to the steady-state value of the flow. A valve-motor control system is studied by linear analysis and the frequency-response characteristic obtained. The transfer function of this system is of the third order if the servomotor load is a mass. Expressions for the damping ratio, natural frequency, and sensitivity are obtained.

THE growing importance of hydraulics is recognized as one of the more significant trends in the engineering of feedback controls. The underlying reasons for the wider acceptance of hydraulics in servomechanisms are (1) the larger power amplification obtainable; (2) the speed of response that can be obtained with these mechanisms; and (3) the compactness of hydraulic power actuators. It has been difficult to predict accurately the dynamic performance of hydraulic systems because of the nonlinear relation between pressure and flow operating in these systems. It is possible to "linearize" many systems of practical interest, however, by considering small motions and applying perturbation theory. Frequency-response analysis techniques then may be applied to determine the dynamic performance of these systems.[2] While situations always exist in which the nonlinear behavior of the system is so pronounced that linear analysis is inapplicable, nevertheless practice has verified that studying the "small-motion" case by means of linear analysis is often very helpful. The method of linearizing hydraulic systems is illustrated by studying in some detail the operation of a specific system commonly encountered. Similar approaches can be employed to study other systems.

Valve-Motor System

A typical feedback control system employing hydraulic elements is illustrated in Fig. 1. Here a hydraulic motor of the fixed-displacement type is controlling the position of a load. Under appropriate circumstances, a cylinder-piston type of hydraulic actuator may be used instead of the rotational motor illustrated. The motor is controlled by a valve supplied with a nominally constant hydraulic pressure by a suitable hydraulic pump. The valve is operated by an electric motor which is driven by an amplifier. The signal to the amplifier is the difference between the required position of the load and its actual position as monitored by the load synchro. Since the input to the amplifier is the error or difference between the input position and the output position, the system is a servomechanism in the accepted sense.

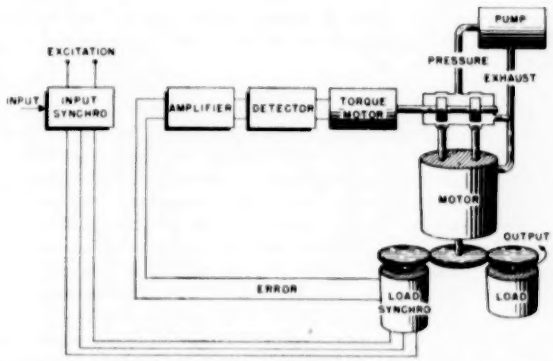


FIG. 1 TYPICAL SERVOMECHANISM EMPLOYING HYDRAULIC ELEMENTS

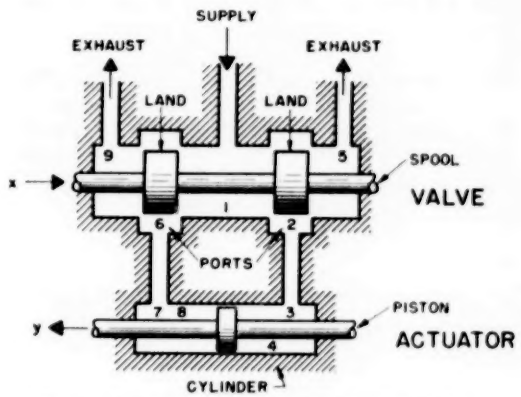


FIG. 2 DIAGRAM OF VALVE AND HYDRAULIC MOTOR

This paper is limited to a discussion of the dynamic properties of the valve-and-motor combination in this servomechanism. Essential elements of the hydraulic valve and motor are shown in Fig. 2. The valve is assumed to be of the "closed-center" type in order that substantial pressure may be supplied to each side of the hydraulic motor in the neutral position of the valve. The motor shown in Fig. 2 is a cylinder-piston type of actuator. The analysis given, however, applies also to the rotational type of motor shown in Fig. 1.

The stations of interest in the valve and motor are indicated by numbers in the diagram, Fig. 2, and are thus referred to in the subsequent discussion. When the valve spool is in its neutral position, the pressure under steady-state conditions at point 3 is

---

[1] Technical Director, Bendix Aviation Corporation, Research Laboratories.

[2] The first instance, to the author's knowledge, of the linearization of hydraulic systems for the purposes of dynamic analysis was made by N. B. Nichols in an unpublished memorandum at the Massachusetts Institute of Technology in June, 1942. Much of the present paper stems from a memorandum by the author written at the Dynamic Analysis and Control Laboratory at M.I.T. in February, 1947.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 22, 1953. Paper No. 53—A-16.

equal to that at point 7. The magnitude of this pressure is a function of the leakage flow which occurs between points 1 and 2, 2 and 5, and similarly between points 1 and 6, 6 and 9.

As the valve spool is displaced to the right, the size of the opening between points 1 and 2 increases as does the opening between points 6 and 9. The hydraulic resistance therefore is reduced between these points and, consequently, the pressure increases on the right side of the motor and decreases on the left side of the motor. Similarly, if the valve spool is displaced to the left of its neutral position, the flow resistance between points 1 and 6 is reduced as well as that between stations 2 and 5 and the opposite action occurs.

In practice, the effective width of the land may be less than, equal to, or greater than the effective width of the ports. These three situations are known as the underlapped, line-to-line, and overlapped valve, respectively. In each case leakage will occur and indeed is very important since it permits a definite pressure to exist on each side of the motor in the rest position.

In the analysis presented, the symbols P and Q are used to represent, respectively, hydraulic pressure and flow. The unit of pressure is the pound per square inch, and that of flow is the cubic inch per second. The symbol R is employed to represent flow resistance. Capital letters with an asterisk superscript indicate the instantaneous value of a quantity; capital letters with no superscript represent the steady-state value; and lowercase letters are employed to represent changes or perturbations in a quantity. Numeral subscripts to a symbol for a variable are used to indicate the station in Fig. 2 at which the quantity pertains.

Thus

P_{12} = steady-state pressure drop between point 1 (supply point) and point 2 (pressure in line to motor)

P_{12}^* = instantaneous pressure drop between the same points

p_{12} = perturbation in pressure drop between points 1 and 2

Obviously

$$P_{12}^* = P_{12} + p_{12} \quad [1]$$

Similarly

Q_{13}^* = instantaneous flow between points 1 and 2, and so on.

The following fundamental relation is assumed between flow and pressure

$$\sqrt{P^*} = R^* Q^* \quad [2]$$

The hydraulic resistance R^* is assumed to be fixed for a particular geometry, and independent of the flow Q^*. The first step in the analysis consists of writing pressure-drop equations and flow equations relating to the individual stations in the valve and motor. For example, the following equations define the flow about the right-hand land and port of the valve

$$\sqrt{P_{13}^*} = R_{13}^* Q_{13}^* \quad [3]$$

$$\sqrt{P_{25}^*} = R_{25}^* Q_{25}^* \quad [4]$$

Similar equations can be written relating pressure, flow, and hydraulic resistance for the left-hand land and port in Fig. 2.

If compressibility effects within the valve itself are neglected, equations of the type of the following may be written

$$Q_{12}^* = Q_{23}^* + Q_{25}^* \quad [5]$$

which states simply that the flow to the right side of the motor is the flow from station 1 to station 2 less the leakage flow from station 2 to station 5 in the valve.

The pressure drop in the lines connecting the valve to the motor often can be neglected. The speed of response of the system is improved if these lines are short and, in practice, the pressure drop in the line is small compared to the pressure drop in the valve itself. The volume of fluid in the line is generally important, however, because of compressibility effects. It is convenient, therefore, to neglect the pressure drop in the line resulting from flow and, in considering compressibility effects, to add the volume of fluid in the line to the corresponding volume in the motor.

The fluid flow to the motor comprises three components, i.e., the flow due to the motion of the driving member (the rectilinear motion of the motor piston or the rotation of the type of motor shown in Fig. 1), the flow due to oil leakage by-passing the motor piston, and the flow due to the compressibility of the oil within the motor or attached line. Thus it is possible to write the following equation

$$Q_{23}^* = Q_{c4}^* + Q_d^* + Q_L^* \quad [6]$$

in which

Q_d^* = flow due to leakage

Q_L^* = flow due to motion of driving member

Q_{c4}^* = flow due to compressibility of oil in chamber 4 and its connecting line

A similar equation can be written for the second side of the motor.

The compressibility component of flow is related to the pressure in chamber 4 by the following expression

$$P_4^* = S_4 \int Q_{c4}^* dt \quad [7]$$

The term S_4 may be called the "hydraulic susceptance." It is related to the compressibility coefficient σ and the volume V_4 of chamber 4 and its attached line by the following

$$S_4 = \frac{1}{\sigma V_4} \quad [8]$$

A similar equation may be written for chamber 8 of the motor. A representative value of σ is 5×10^{-6} lb^{-1} in^2.

The pressure effective in moving the load is given by

$$P_L^* = P_4^* - P_8^* \quad [9]$$

It is convenient to define the hydraulic impedance of the load as the following ratio

$$Z_L = \frac{P_L^*}{Q_L^*} \quad [10]$$

Linearization of Analysis

Equations of the type of Equations [3] to [10] define the valve-motor combination. Since the pressure-flow equations are nonlinear, however, a rigorous analysis is generally not practical. While the equations as written cannot be handled directly by linear analysis, a linear approximation to the equation can be developed which is satisfactory for many purposes. The procedure followed is to assume that each of the terms in Equation [2] is the sum of steady-state and perturbation components, and that the latter is small relative to the former. A set of linear equations approximately correct for the perturbation components may then be derived. While these equations will be strictly true for infinitesimal perturbations only, they will apply, in so far as results of engineering accuracy are concerned, to changes of appreciable magnitude.

An equation approximately true for small perturbations of the fundamental pressure-flow relationship can be found by taking differentials of Equation [2]. The resulting equation is

$$Rq + Qr = \frac{1}{2}\frac{p}{\sqrt{P}} \quad \text{[11]}$$

As stated previously, capital letters in Equation [11] indicate the steady-state values and lower-case letters indicate perturbations. The instantaneous value of each quantity is given by the sum of the steady-state and perturbation terms. Equation [1] is a typical relation. A change in the flow resistance R^* is due to a displacement of the valve spool relative to the ports.

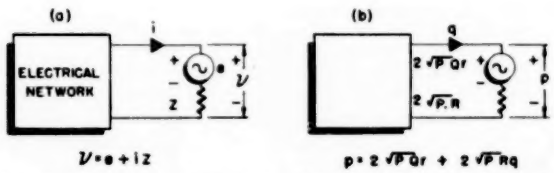


FIG. 3 ELECTRICAL ANALOG OF VARIABLE HYDRAULIC ORIFICE

An electrical network representation of Equation [1] is useful. If, as shown in Fig. 3(*a*), an electrical network supplies current to a series combination of a generator of voltage e and a resistance of impedance Z, the voltage drop across the series combination is given by the expression shown. Now if voltage and current are made analogous to the pressure change p, and flow change q, respectively, in a hydraulic system, the electric circuit shown in Fig. 3(*b*) represents Equation [11]. A change in hydraulic resistance is similar in its effect to a generator whose electromotive force is proportional to the change in resistance, in series with an electrical resistive element proportional to the steady-state value of the hydraulic resistance.

ELECTRICAL ANALOGS

An electrical analogy of the complete valve-motor system can be drawn, following the procedure outlined and still assuming small perturbations. The resulting network is shown in Fig. 4. Here, voltage is directly analogous to the change in pressure drop p, and current is analogous to the change in fluid flow q. The electromotive force of generator A represents the change in hydraulic resistance between points 1 and 2 in Fig. 2 caused by a motion of the valve, and resistance B represents the average hydraulic resistance between these two stations. Similarly, generator D represents the change in hydraulic resistance between stations 2 and 5 and electrical resistance C represents the average hydraulic resistance between these stations. The generators and resistances in the lower half of the circuit relate to the left side of the valve. The circuit presumes there is no pressure differential between stations 5 and 9.

The susceptances (inverse capacitances) S_4 and S_8 are the electrical representation of the oil compressibility effects in the hydraulic lines and the motor. The current through an electrical susceptance is related to the voltage across it by the following expression

$$e = S \int i dt \quad \text{[12]}$$

The similarity of this equation to Equation [7], expressing the oil compressibility effect, establishes the electrical analogy.

It is emphasized that the circuit in Fig. 4 is useful only for determining dynamic conditions and does not apply to determining the steady-state values of pressure and flow existing in the system. While studying a hydraulic system by making use of an electric equivalent circuit is not necessary, the familiarity of all engineers with simple electric circuits makes this a convenience and permits easy visualization of many of the significant dynamic effects.

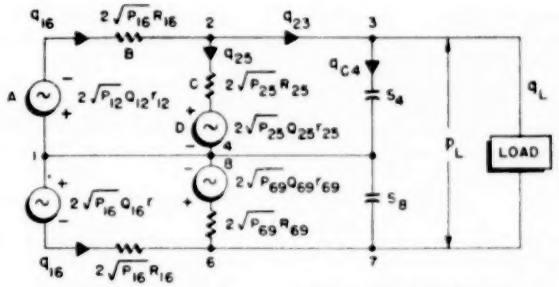


FIG. 4 ELECTRICAL ANALOG OF HYDRAULIC-VALVE MOTOR

Some sacrifice in generality must be made to simplify further the equivalent circuit in Fig. 4. The following assumptions are not restrictive for many applications, and in cases where they do not hold, the problem may be handled by dealing with the general equivalent circuit in Fig. 4: (1) It is assumed that the problem of principal interest is that of determining the response of the motor to a displacement of the valve from its neutral position. (2) The valve is assumed to be symmetrical. (3) It is assumed that the oil volumes between the motor piston and valve are equal for each side of the motor. If the motor is a rotational type, the third condition is not particularly restrictive; however, if the motor is an actuator with long travel and if the volume of oil in the lines is small, this supposition is not valid. (4) All increments in valve resistance are assumed to be related to the valve displacement by the same proportionality factor.

The foregoing assumptions lead to the following equations

$$P_{12} = P_{25} = P_{16} = P_{69} = \frac{P_s}{2} \quad \text{[13]}$$

$$R_{12} = R_{25} = R_{16} = R_{69} = R_v \quad \text{[14]}$$

$$Q_{12} = Q_{25} = Q_{16} = Q_{69} = \frac{1}{2}Q_0 \quad \text{[15]}$$

$$S_4 = S_8 = \frac{1}{2}S_m \quad \text{[16]}$$

$$r_{12} = r_{69} = -k_{rx}x \quad \text{[17]}$$

$$r_{25} = r_{16} = +k_{rx}x \quad \text{[18]}$$

The quantity P_s in Equation [13] is the supply pressure. In Equation [15] the quantity Q_0 is the total flow from the supply to the valve when the servomotor is not operating and leakage across the servomotor is negligible. The variable x is the mechanical displacement of the valve from its neutral position, measured in inches. The coefficient k_{rx} is the proportionality factor relating a change in the hydraulic resistance in the valve port with the displacement of the valve. Its value is determined in a following section.

If V_m is the volume of oil under compression on one side of the motor, including the volume of the pipe connecting one side of the valve to the motor, then

$$S_4 \text{ or } S_8 = \frac{1}{\sigma V_m} \quad \text{[19a]}$$

and

$$S_m = \frac{2}{\sigma V_m} \quad \text{[19b]}$$

A simplified electric equivalent circuit can be derived by mak-

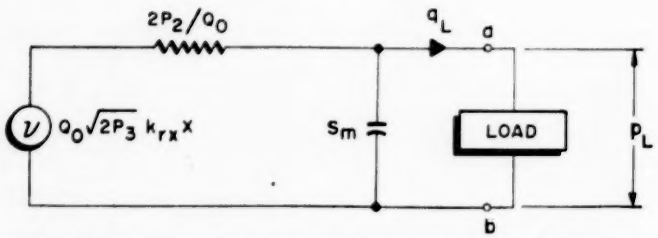


FIG. 5 ELECTRICAL ANALOG OF SYMMETRICAL VALVE-MOTOR SYSTEM

ing use of the foregoing assumptions and is shown in Fig. 5. The electromotive force $Q_0\sqrt{2P_s}\,k_{rx}x$ is the net driving force acting upon the system due to the displacement of the valve from its neutral position. The ratio $2P_s/Q_0$ is the effective internal hydraulic resistance of the valve. Compressibility effects are represented by the susceptance S_m.

The quantity p_L is the hydraulic pressure that actually operates on the load and the quantity q_L is the fluid flow in the motor resulting from the displacement of the load. The hydraulic quantities p_L and q_L are related directly to mechanical quantities by

$$\text{Force or torque} = Ap_L \quad [20]$$

$$\text{Rectilinear or angular velocity} = q_L/A \quad [21]$$

If the hydraulic motor is a cylinder and piston, the coefficient A is the effective area of the piston. On the other hand, if the hydraulic motor is a rotational one, the coefficient A is an equivalent area multiplied by an effective radius arm.

Often the mechanical load of the servomotor comprises mass only. In such cases one of the following sets of equations may be written

$$F = M\frac{d^2y}{dt^2} \quad [22a]$$

for rectilinear motors, or

$$T = J\frac{d^2y}{dt^2} \quad [22b]$$

for rotational motor, where

F = force developed by motor, or
T = moment developed by motor
M = mass of load, or
J = moment of inertia of load
y = translation or rotation

The relations given in Equations [20] and [21], relating hydraulic and mechanical quantities, may be substituted in Equations [22].

For a rectilinear motor

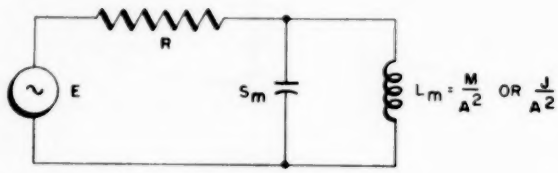
$$p_L = \frac{M}{A^2}\frac{dq_L}{dt} \quad [23a]$$

and for a rotational motor

$$p_L = \frac{J}{A^2}\frac{dq_L}{dt} \quad [23b]$$

The electrical analog of a mass load is determined quickly. The relationship between voltage and current in an inductive electrical element is given by the following expression

$$e = L\frac{di}{dt} \quad [24]$$



$E = \sqrt{2P_s}\,Q_0\,k_{rx}\,x$
$R = 2P_s/Q_0$
$S_m = 2/\sigma V_m$

FIG. 6 EQUIVALENT NETWORK FOR VALVE MOTOR WITH MASS LOAD

Equation [24] is analogous to Equation [23], thus establishing that in Fig. 5 the electrical element required to represent a mass load to the servomotor is an inductance. Thus, for the important case in which the load is primarily a mass, the equivalent circuit of the hydraulic valve-motor system may be represented by Fig. 6.

FREQUENCY RESPONSE OF VALVE MOTOR

The frequency-response relation between the output motion of the motor and displacement x of the valve now can be computed for the case of the mass load represented in Fig. 6. The resulting expression is

$$\frac{y(j\omega t)}{x(j\omega t)} = \frac{Q_0^2 k_{rx}}{\sqrt{2P_s}\,A}\,\frac{1}{j\omega}\,\frac{1}{\dfrac{(j\omega)^2}{\omega_0^2} + \dfrac{2\zeta j\omega}{\omega_0} + 1} \quad [25]$$

The form of Equation [25] indicates that in the steady state, the output velocity is proportional to valve displacement, and that the transient relation between output velocity and input displacement is the familiar second-order function in which the undamped natural frequency in radians per second is given by

$$\omega_0 = \sqrt{\frac{S_m}{L_m}} \quad [26]$$

The damping ratio ζ is

$$\zeta = \frac{Q_0}{4P_s}\sqrt{L_m S_m} \quad [27]$$

The natural frequency f_0, in cycles per second, can be written as follows

$$f_0 = \frac{A}{\pi}\sqrt{\frac{1}{2\sigma V_m M}} \quad [28]$$

It is particularly convenient to use Equation [28] if the servomotor is of the rectilinear type. This expression gives the natural frequency of the system in terms of the output mass of the servo-

motor M, the effective area of the piston in the actuator A, the compressibility factor σ, and the oil under compression V_m.

If the servomotor is a rotational one, it is convenient to express the natural frequency of the system as follows

$$f_0 = \frac{d}{2\pi^2}\sqrt{\frac{1}{2\sigma V_m J}} \quad [29]$$

The quantity d is defined as the "displacement" of the rotational motor and may be expressed in cubic inches per revolution.

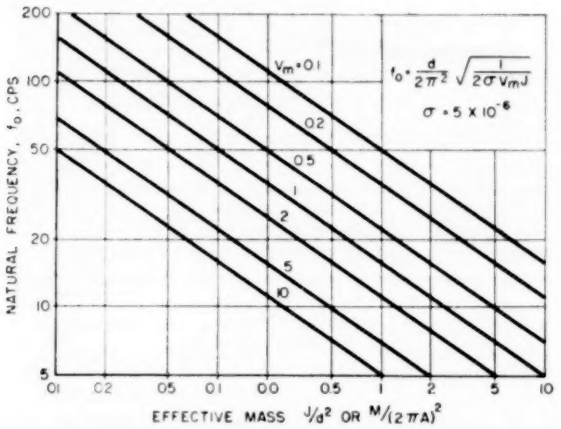


FIG. 7 UNDAMPED NATURAL FREQUENCY OF VALVE MOTOR WITH MASS LOAD

The natural frequency of a valve-motor combination with mass load is plotted in Fig. 7. For the rotational motor, the abscissa is the ratio of the moment of inertia of the load to the square of the displacement of the motor, the quantity J/d^2. The moment of inertia is expressed in pound-inch-second^2, and the displacement in cubic inches per revolution.

For a rectilinear motor, the abscissa in Fig. 7 is the quantity

$$\frac{M}{(2\pi A)^2}$$

in which M is the mass of the moving member of the servomotor and its attached load expressed in pound-inch-second,^{2} and A is the effective area of the ram expressed in square inches. In each case the ordinate is given in cycles per second and the parameter V_m is the volume of oil under compression between one side of the motor and the valve, measured in cubic inches.

The equivalent circuit of Fig. 5 and the expression for frequency response, Equation [25], contain elements that remain to be related to the valve geometry. Both the system sensitivity factor and the damping ratio are related to the leakage properties of the valve, and an estimate can be made of these factors, although it should be emphasized that in conventional systems the accuracy with which these two factors can be calculated is considerably less than that with which the system natural frequency can be computed.

The oil flow through the valve may be estimated by considering the port as an orifice. The flow of oil through an orifice is given by the following expression

$$Q^* = C_0 A_0 \sqrt{\frac{2g}{\rho}} P^* \quad [30]$$

Here, P^* and Q^* are, respectively, the pressure drop across, and the flow through the orifice. A_0 is the area of the orifice, C_0 is an orifice coefficient, and g and ρ are, respectively, the acceleration of gravity and the density of the fluid. The orifice coefficient C_0, unfortunately, is not constant; however, a representative value of 0.6 gives useful results. If Equation [2], the flow relation that has been used throughout this analysis, is compared with Equation [30], it may be seen that the hydraulic resistance is given by

$$R^* = \frac{1}{C_0 A_0}\sqrt{\frac{\rho}{2g}} \quad [31]$$

For an underlapped type of valve, the geometrical relation between land and port can be represented approximately by Fig. 8. In this figure, A, B, C, and D indicate the outline of the port. The area A, E, F, D may be considered the opening between stations 1 and 2 in Fig. 2, while the area H, G, C, B is the opening between stations 2 and 5. These two areas, which are equal at the neutral position of the symmetrical valve, determine the leakage flow in the valve.

The series resistance, $2P_s/Q_0$ in Fig. 5 is determined by the valve leakage. If the valve is symmetrical and the area of the opening A, E, F, D is denoted as A_0, the effective series resistance is

$$2P_s/Q_0 = \frac{1}{C_0 A_0}\sqrt{\frac{\rho}{g}} P_s \quad [32]$$

For the case in which the load is entirely mass, the damping ratio can be related to the port opening by the following equation

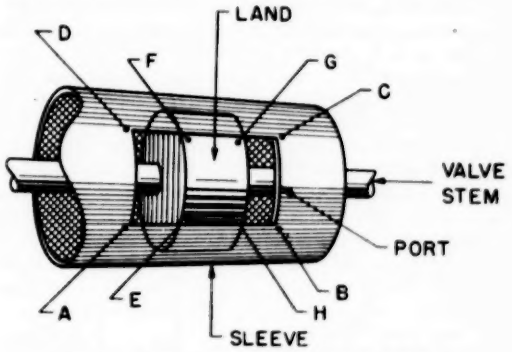


FIG. 8 RELATION OF VALVE LAND AND PORT

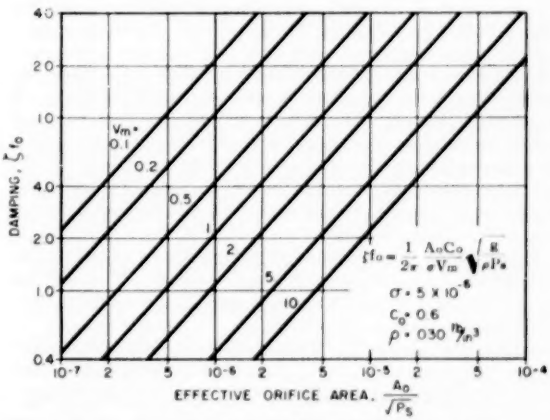


FIG. 9 DAMPING OF VALVE MOTOR WITH MASS LOAD

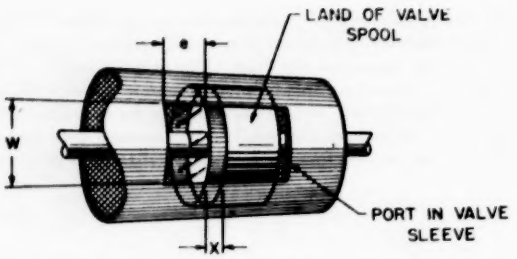


FIG. 10 LAND-PORT GEOMETRY FOR DISPLACED VALVE

$$\zeta f_0 = \frac{1}{2\pi} \frac{A_0 C_0}{\sigma V_m} \sqrt{\frac{g}{\rho P_s}} \quad [33]$$

This relation is plotted in Fig. 9. The natural frequency f_0 may be obtained from Fig. 7.

The proportionality constant k_{rx} relating the change in the hydraulic resistance of the valve with valve displacement also may be determined from geometrical considerations in the valve.

Fig. 10 illustrates the position of an underlapped land, slightly displaced from the neutral position. The orifice area is given by

$$A_x = w(e + x) \quad [34]$$

in which x is the displacement from the neutral position, e is the distance through which the valve must be displaced in the negative direction to close the opening, and w is the effective width of the port. Using Equation [31], the effective resistance of this opening is

$$R_v^* = \frac{1}{w(e + x)C_0} \sqrt{\frac{\rho}{2g}} \quad [35]$$

The proportionality k_{rx} may be defined as

$$k_{rx} = \left. \frac{dR_v^*}{dx} \right|_{x=0} \quad [36]$$

Therefore

$$k_{rx} = \frac{1}{eA_0C_0} \sqrt{\frac{\rho}{2g}} \quad [37]$$

The sensitivity factor of Equation [25] may now be computed

$$\frac{Q_0^2 k_{rx}}{\sqrt{2P_s A}} = \frac{2C_0 A_0}{eA} \sqrt{\frac{g}{\rho} P_s} \quad [38]$$

It should be emphasized that the computation of the sensitivity factor from the valve geometry is often of limited usefulness. It is generally desirable for other reasons to construct valves in which the underlap e is as close to zero as practical. Under these circumstances, the effect of diametral clearance on leakage and valve sensitivity cannot be neglected. Also, the leakage in the motor, while contributing in a nonlinear fashion, definitely increases the damping ratio.

The following is an example of the calculation of the natural frequency and damping of a valve-motor system. The pertinent characteristics of a typical hydraulic servomotor useful for control applications are given in Table 1.

Assume the load to equal 0.1 lb-in-sec^2, referred to the motor shaft. With such a load the moment of inertia of the motor itself is insignificant, and this is generally true. Since the displacement d is 0.25 cu in., the ratio J/d^2 is 1.6.

The volume of oil under compression in the motor is only 0.4 cu in. However, to this must be added the corresponding volume of oil in the lines and valve. If 1 cu in. is taken as a typical value, the quantity V_m is 1.4. Referring to Fig. 7, it is found that the undamped natural frequency of the valve-motor is approximately 11 cps.

If the effective port width w is $\frac{1}{2}$ in. and the underlap e is 0.0005 in., the orifice valve area A_0 is 2.5×10^{-4} sq in. If the operating pressure P_s is 2500 psi, the value of $A_0/\sqrt{P_s}$ is 5×10^{-6}. Referring to Fig. 9, and again letting $V_m = 1.4$ cu in., it may be seen that ζf_0 is approximately 8.0. Since f_0 has been determined to be 11 cps, the damping ratio ζ is approximately 0.7.

TABLE 1 MOTOR CHARACTERISTICS

Maximum pressure, psi	3000
Maximum speed, rpm	4500
Displacement (d), in^3/rev	0.25
Horsepower (maximum)	5.0
Moment of inertia, lb-in-sec^2	5×10^{-4}
Volume of oil under compression, in^3	0.4
Volumetric efficiency per cent	95
Torque/inertia (maximum)	200,000

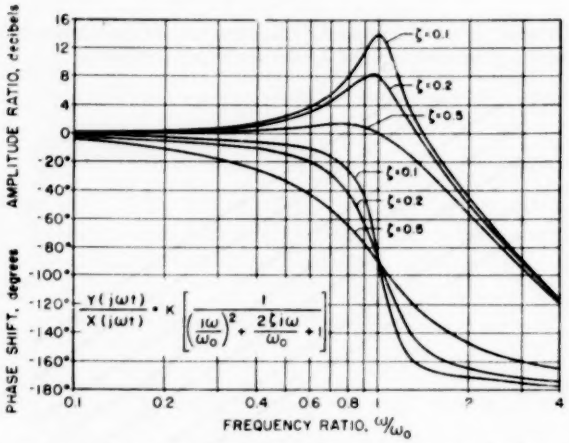


FIG. 11 FREQUENCY RESPONSE OF TYPICAL HYDRAULIC SYSTEMS

The frequency responses of typical hydraulic systems are shown in Fig. 11. It should be noted that this figure gives the amplitude and phase relation between the displacement X of the valve and $\dot{Y}$ the "velocity," or time rate of change, of the servomotor output. It is a characteristic of hydraulic systems that the damping ratio is generally less than unity since for efficiency and other reasons it is desirable to keep the leakage in the valve and motor small. It is almost always desirable to operate the system with small hydraulic damping and if the resulting peak in the amplitude response is objectionable, either to eliminate it by compensation in the feedback system or to so increase the natural frequency of the hydraulic system that this peak lies well outside the pass band of the over-all system.

ELECTRICAL ANALOGS FOR SERVO LOADS

The foregoing analysis holds when the load on the hydraulic motor is a mass element, and it has been shown that the equivalent electrical element representing the load is an inductance. Often the load on the servomotor comprises other elements and in these cases other equivalent circuits can be found. A few such circuits for commonly encountered loads are shown in Fig. 12.

Fig. 12(a) represents the case already considered and shows the inductive element to be connected between points a and b in Fig. 5, to represent a mass load. In Fig. 12(b) the load comprises three elements, i.e., a mass M; viscous friction, f_v, expressed

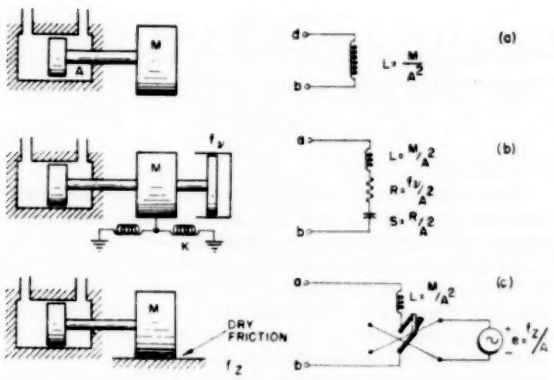


FIG. 12 ELECTRICAL ANALOGS OF TYPICAL SERVO LOADS

in pounds-inch-second^{-1}; and a spring K, expressed in pounds-inch^{-1}. This type of load often arises in an aircraft-control application in which the spring load K represents the aerodynamic force acting upon the control surfaces of the aircraft. The electrical analog for such a load is shown to be a resistance-capacitance-inductance series circuit.

Another commonly encountered load for a hydraulic servomotor is dry or "coulomb" friction. This friction element and its analog is illustrated in Fig. 12(c). Coulomb friction always acts in a direction to oppose the motion of the piston, and its magnitude is independent of the velocity of the piston. While this load is nonlinear, it still may be represented by an electrical element and useful results may be obtained by appreciating how such a load affects the equivalent electric circuit. For dry friction, f_z is represented in the electric circuit by a direct voltage source. The voltage source must be considered as instantaneously reversible and always acting in a direction to oppose the flow of current through the load branch of the equivalent circuit.

CONCLUSION

The method outlined in this paper by means of which the characteristics of the hydraulic system are linearized and its response studied by means of frequency-response techniques has been applied to the development of a number of hydraulic control systems during the past 6 years. The limitations of the method must not be forgotten. However, this method of analysis has been instrumental in improving the dynamic performance of hydraulic servosystems by a considerable factor. The result has been that servosystems with natural frequencies of better than 100 cps have proved practical and successful.

Discussion

R. H. MACMILLAN.[3] The method of linearization used by the author for the valve characteristic is also successfully employed by Harpur[4] in a valuable paper on flying-control servos. It may be of interest to note that the possibility of using a valve with large overlap, which operates solely by variation of leakage path, so that the valve ports are never allowed to break open, is being investigated in Britain. Such a system, although still nonlinear, appears to possess manufacturing and operational advantages in certain applications.

J. L. BOWER.[5] Quite evidently this paper is a digest of much experience in the field of throttling servo valves and related devices. It presents a clear summary of the principal parts of the theory in a form most useful to the practicing servo engineer. It would seem, however, that the practical importance of hydraulic reactions resulting from the flow of oil would demand their mention in some form. These effects are notoriously the cause of many troublesome oscillations in valve systems, and the reason why great effort has been spent in the past on compensated valves. Since much of that effort was applied in laboratories with which the author was formerly associated, the effects certainly hold no mysteries for him. While the relations presented, being in terms of valve displacement, rather than force, are correct, it would seem better to warn the unwary reader that the reactions exist and are ignored for a stated reason. The relations as they stand in the paper are practically usable, without attention to reaction forces, in case an overpowerful torque motor and amplifier and a stiff centering spring are provided. They also may be used when the displacement of the stem is measured and adequately controlled by feedback.

Two other effects that probably deserve mention are (1) the effect of the entrained air on oil compressibility; and (2) the fact that a geometrically symmetrical valve may not, and generally will not, yield a symmetrical diagram in Fig. 4 of the paper for large displacements unless the center pair of ports has displacement characteristics identical with the outer pair of ports. The latter condition represents more than simple symmetry.

J. L. SHEARER.[6] Much of the recent work that has been done on the development of valve-controlled hydraulic systems stems from the early technical guidance provided by the author in the development of high-performance hydraulic systems in the Dynamic Analysis and Control Laboratory at the Massachusetts Institute of Technology. A thorough understanding of the factors which influence the dynamic performance of hydraulic systems explains why valve-controlled systems are finding wider use in military and industrial applications and justifies the intensive research and development work that has been done in this field.

It is true that a control valve can maintain effective control of a system only when it dissipates part of the power carried by the fluid flowing through it, and as a result the over-all mechanical efficiency of a valve-controlled system is usually inferior to a variable-displacement-pump controlled system. However, the designer usually is faced with other considerations such as simplicity, cost, and space requirements in addition to speed of response.

Frequency-response techniques have proved to be very useful in the evaluation of real systems in the laboratory. In this connection it is interesting to note that laboratory measurements using small-amplitude inputs do very often agree well with predicted performance based on linearized analyses. Although the author states that a linearized analysis of a control valve can hold only when the change in flow is small compared to the initial flow, recent analytical and experimental work indicates that it is much more important that the change in pressure drop across the valve be small compared to the initial pressure drop across the valve.[7]

It is interesting to note that when the load comprises only mass,

[3] University Lecturer, University of Cambridge, Cambridge, England.

[4] "Some Design Considerations of Hydraulic Servos of Jack Type," by N. F. Harpur, Proceedings of the Institution of Mechanical Engineers, London, England, 1953.

[5] Engineering Specialist—Research, North American Aviation, Inc., Downey, Calif.

[6] Assistant Professor of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[7] "Pressure-Flow Relationships of 4-Way Valves," by J. F. Blackburn, Trans. ASME, vol. 75, 1953, pp. 1163–1169; and "Dynamic Characteristics of Valve-Controlled Hydraulic Servomotors," by J. L. Shearer, Trans. ASME, vol. 76, 1954, pp. 895–903.

system damping must come from valve underlap (and therefore quiescent leakage through the valve) or motor leakage. When it is necessary to keep the quiescent-leakage flow through the valve to a minimum, leakage across the motor or viscous friction on the output shaft must be introduced to provide necessary damping.[7]

The author is to be commended for his early recognition of the merits of valve-controlled systems and the strong impetus that he has given to their development.

HAROLD GOLD.[8] The usefulness of the technique of electrical analogy in the analysis of the dynamic response of hydraulic circuits is clearly demonstrated by the readily applied charts prepared by the author. An additional advantage of the technique that is perhaps not apparent lies in the ease with which additional factors can be taken into account once a basic equivalent network has been derived. Therefore the author's analysis can be extended to include the effect of factors not specifically mentioned in the paper. In this instance the writer will employ the author's equivalent network (Fig. 6) to treat the effect of linear (zero-order) feedback and the effect of oil mass.

Based on the assumptions of small flow changes and of equal flow and equal pressure drop across each of the four valve ports, the pressure drop across each valve port is

$$P = \frac{(Q_0/2)^2}{(Da)^2} \quad [39]$$

where D = dimensional constant in orifice equation, a = open area of valve port.

Equation [39] differentiated with respect to a is

$$\frac{dP}{da} = -2\left(\frac{Q_0/2}{D}\right)^2 \frac{1}{a^3} \quad [40]$$

For small movements of the valve around the neutral position the change in pressure drop across a valve port is

$$\Delta P = \frac{dP}{da}\,\Delta a \quad [41]$$

From geometric considerations and Equations [40] and [41], the change in pressure difference across the piston is

$$\Delta P_P = 4\left(\frac{Q_0/2}{D}\right)^2 \frac{\Delta a}{a^3} \quad [42]$$

In the case of linear feedback

$$\Delta a = rwy \quad [43]$$

where f = feedback ratio (ratio of valve travel to piston travel at fixed input); y = position error of piston.

From the equilibrium conditions

$$a = \frac{Q_0}{2}\,\frac{1}{D\sqrt{\frac{P_s}{2}}} \quad [44]$$

Equations [43] and [44] substituted in Equation [42] yield the proportionality between the opposing pressure difference set up across the piston and the deviation of the piston from the neutral position

$$\Delta P_P = \frac{\sqrt{8}DP_s^{3/2}\,rwy}{Q_0} \quad [45]$$

The volume swept by the piston in each half of the cylinder for a displacement y is

$$V_P = A_P y \quad [46]$$

The ratio $\Delta P_P/V_P$ is equivalent to an electrical susceptance. The equivalent susceptance due to feedback is then

$$S_f = \frac{\sqrt{8}DP_s^{3/2}\,rw}{A_P Q_0} \quad [47]$$

The equivalent network of the servomotor under a mass load is then the circuit of Fig. 6 of the paper with the addition of a susceptance S_f in series with the load inductance.

The dynamic characteristics of the modified network are subject to a wide variety of possibilities if no limits are imposed upon the values of the circuit components. In the present application, however, consideration of the probable range of values of the components permits a simplified but useful evaluation of network dynamics. The first limitation that can be imposed is that a practical servomotor must exhibit a reasonably well-damped response. This limitation combined with an evaluation of probable values of L_m and S_m permits the reduction of the network to one in which the effect of S_m can be neglected. In this case the dynamic characteristics of the network can be described by a damping ratio which is

$$\zeta = \frac{P_s}{Q_0\sqrt{L_m S_f}} \quad [48]$$

A convenient form of Equation [48] is obtained by introducing the quantity "no-load time constant." This characteristic response parameter can be expressed as follows

$$T = \frac{\sqrt{2}\,A_p}{Drw\sqrt{P_s}} \quad [49]$$

Equations [47] and [49] substituted in Equation [48] yield

$$\zeta = \frac{A_p}{2}\sqrt{\frac{P_s T}{Q_0 M}} \quad [50]$$

Equation [50] indicates that the response of the servomotor around the neutral position becomes less damped as the speed of response is increased (T reduced). In the case of servomotors employing electrical valve actuators (such as shown in Fig. 1 of the paper), Equation [50] may be applied if the characteristic time constant of the actuator is short compared with the characteristic time constant of the servomotor. In this configuration the no-load time constant is not fixed in a given design but is an inverse function of the gain of the amplifier. Thus the maximum allowable amplifier gain under which the servomotor can be operated is extended by a reduction in valve leakage or a reduction in load mass.

In the case of lightly loaded servomotors, the effective oil mass in the lines between the valve and piston may add significantly to the load mass. By treating the oil mass as an inductance in series with the load inductance the effect of oil mass is readily taken into account.

The equivalent inductance of a line may be written

$$L_c = \frac{4\rho l}{\pi d^2} = \frac{\rho l^2}{V} \quad [51]$$

where l = length of line; d = diameter of line; V = volume of line.

It can be seen readily from Equation [51] that a reduction in volume of oil in the connecting lines, that is obtained by a reduction in tube diameter only, increases the line inductance and hence also increases the total inductive load on the servomotor piston.

[8] Aeronautical Research Scientist, Lewis Flight Propulsion Laboratory, National Advisory Committee for Aeronautics, Cleveland, Ohio.

Synthesis of Optimum Feedback Systems Satisfying a Power Limitation

By J. H. WESTCOTT,[1] LONDON, ENGLAND

The paper describes two methods of synthesizing feedback systems in which stability considerations are qualified by an overriding limit on power demand. The three basic requirements for either method are (*a*) a spectral density function of input signals, (*b*) a transfer function of fixed elements (i.e., basic elements essential for the performance of the task), and (*c*) the maximum rating of the power source. Examples are given. In one method the form of the stabilizing means has to be specified. The criterion, that the mean-square error be a minimum subject to the power limitation, is applied. This gives as a result the optimum design parameters of the specified stabilizing means. The other method, based upon the same criterion, uses a variational principle. A solution by this method gives a stabilizing means for the system which, within the specified power limit, cannot be improved. One example of this method illustrates the desirability of allowing a finite time delay between the input and the output of a feedback system.

Introduction

THE classical feedback-system design methods are firmly based on considerations of stability. In order to make an analysis of a closed-loop system of a practical form it is necessary to make the assumption that superposition of responses applies, at least over a limited range. A complete analysis without this assumption becomes impossibly difficult and little progress can be made toward a logical design even with relatively simple systems. With this in mind it is proposed to introduce here linear methods of analysis that make the most of a difficult situation by placing a restriction upon the mean power that can be demanded under representative operating conditions. In this fashion the consideration of stability in a system is qualified by an overriding limitation in the power that may be demanded.

Feedback systems are essentially devices for controlling a source of power but present methods pay little attention to the power limitation of that source. Nevertheless the size of the source is the basic economic factor in the design of the system. The design of a high-quality control system can be completely frustrated in its practical form because impossible demands are made of the power source provided. The methods described enable the best system to be synthesized for a specified power demand. The accuracy of a feedback system depends to some degree on the subtlety of its design, but it is fundamentally limited in a manner that can be assessed when a representative cross section of signals is known and the nature of the load and limitation of the power source are specified. These three factors form the basic elements of a design. The specification of the power source and load together represent the "fixed elements" of the system which have to be accepted in the design. It then remains to find the arrangement giving the highest performance accuracy. The design methods to be described specify complementing elements and connections which, in conjunction with the fixed elements, give the minimum mean square error within the power limit specified.

Two distinct methods of design are considered. One is a simplified method in which it is necessary at the outset to decide the interconnection of the elements to be employed. The procedure then consists in varying the parameter values that are not specified previously while maintaining the power demand constant until the system gives minimum mean-square error. The resulting system is referred to as parameter-optimized. Its success depends to some extent on the wisdom of the initial choice of interconnections and elements. There always must remain some doubt as to whether some other interconnection might not lead to a better result. The other design method overcomes this objection completely by relieving the designer of the onus of an initial choice of elements. This more refined method is based upon the calculus of variations and when the fixed elements have been specified allows of no choice in the complementing elements, but gives directly an absolute optimum for the prescribed power which cannot by any linear means be improved upon. These are referred to as variation-optimized systems. Sometimes they may be difficult to realize in a practical form since the method takes no account of practical convenience, but their existence is also of great value as a standard, which cannot be bettered, against which to compare systems of a more conventional type.

In many industrial applications, feedback systems are employed in a role where a pure time delay between input and output would be of no consequence and could be allowed for in the process. The variational method of design is used to show that in such cases a properly chosen time delay can lead to a substantial reduction in the minimum mean-square error attainable for the same input, fixed elements and power limit. The allowing of such a time delay means that notice of sudden change in output is given to the system which then "prepares" for a sudden movement of the output member. In this fashion instantaneous "error" may be quite large giving a forcing action, while actual error taking account of the delay is small.

Parameter Optimized System With Power Limit

Procedure. The first requirement as a basis of design is a statistical statement concerning the input signals. This is needed in the form of a spectral density function derived from representative sample inputs. Details of the power source and load give the transfer function of fixed elements. The general form of the stabilizing networks has to be decided. The transfer functions of the system are derived in terms of the parameters of the stabilizing networks and gain factors of the system which constitute the design variables. It is then possible to write down in terms of the design variables an integral consisting of the mean-square error of the system in response to the input, to which is added an expression representing the power demand multiplied by an undefined multiplier ν; thus

$$I = \overline{\epsilon^2} + \nu P \quad \ldots\ldots [1]$$

[1] Lecturer, Electrical Engineering Department, Imperial College of Science and Technology.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 3, 1953. Paper No. 53—A-17.

where

$\overline{\epsilon^2}$ = mean-square error
P = power demand

The integral can be obtained in terms of the algebraic parameters using standard forms.[2] I is now minimized with respect to the design parameters giving a function involving ν, but since the power is to be considered constant the variation is independent of the added constant νP. Thus a given value of ν yields a particular power demand and a corresponding minimum mean-square error. By considering a range of ν-values a graph can be drawn of minimum mean-square error against power, from which the particular value of ν corresponding to the specified power may be selected and the optimum system parameters for this power found by substitution.

Example. A position-control servomechanism is required to follow an input signal consisting of a series of time intervals whose terminal points have a Poisson distribution of shots on a line of mean density $1/T_3$ per sec. During each interval the velocity has a constant value which is independent of the velocity in other intervals; the velocities are randomly distributed with zero mean and a mean-square value $\overline{a^2}$. The spectral density function of such an input signal may be represented,[3] in operational notation, as follows

$$p\theta_i(p) = \frac{(2\overline{a^2}T_3)^{1/2}}{(T_3p + 1)} \quad \ldots\ldots [2]$$

where p is a complex variable.

For the power source of the servomechanism an electric motor

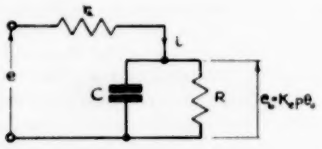


FIG. 1 EQUIVALENT CIRCUIT OF AN ELECTRIC MOTOR

is chosen. In Fig. 1 is shown the equivalent electric circuit of the motor (with constant field) and the load. Capacitance C represents the combined inertia of motor armature and load. Resistance R represents the internal damping of the motor and viscous friction of the load. The voltage e_b across the parallel combination is proportional to the speed of the motor. Resistance r_a is the armature resistance through which the total current flows; armature inductance is neglected. The power has two components; a mechanical power associated with friction losses represented by the dissipation in resistance R, and an electrical power associated with the armature loss in resistance r_a. The mechanical power density is given by

$$\delta P_f = \frac{e_b(p)\, e_b(-p)}{R} \frac{\delta p}{2\pi j}$$

but $e_b(p) = K_e\, p\theta_o(p)$, hence

$$\delta P_f = \frac{K_e^2}{R} [p\theta_o(p)]\ [-p\theta_o(-p)] \frac{\delta p}{2\pi j} \quad \ldots\ldots [3]$$

whence the mean power P_f is[4]

$$P_f = \frac{K_e^2}{R} \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} [p\theta_o(p)]\ [-p\theta_o(-p)]\, dp \quad \ldots\ldots [4]$$

The electrical power density is

$$\delta P_r = i(p)i(-p)\, r_a \frac{\delta p}{2\pi j} = [e_b(p)\, Y_e(p)]\ [e_b(-p)\, Y_e(-p)]\, r_a \frac{\delta p}{2\pi j}$$

where

$$Y_e = Cp + \frac{1}{R} = \frac{T_1p + 1}{R} : T_1 = RC$$

Hence

$$\delta P_r = [e_b(p)\,(T_1p + 1)]\ [e_b(-p)\,(-T_1p + 1)] \frac{r_a}{R^2} \frac{\delta p}{2\pi j}$$

Substituting for e_b

$$\delta P_r = \frac{K_e^2}{R^2}\, r_a [p\theta_o(p)\,(T_1p + 1)]\ [-p\theta_o(p)\,(-T_1p + 1)] \frac{\delta p}{2\pi j} \quad \ldots\ldots [5]$$

whence the mean power P_r is

$$P_r = \frac{K_e^2\, r_a}{R^2} \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} [p\theta_o(p)\,(T_1p + 1)]\ [-p\theta_o(-p)\,(-T_1p + 1)]\, dp \quad \ldots\ldots [6]$$

The amplifier associated with the motor has negligible delay and an adjustable gain such that the over-all combination from error angle ϵ to output shaft has a steady-state transfer $(p\theta_o/\epsilon)$ of K_1 radians per sec/radian. The transfer function to the output shaft of the fixed elements is thus given by

$$F(p) = \frac{K_1}{p(T_1p + 1)} \quad \ldots\ldots [7]$$

where K_1 is variable.

The general arrangement of the system is shown schematically in Fig. 2, in which it will be seen that a stabilizing feedback path

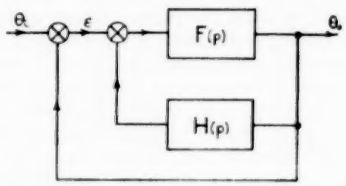


FIG. 2 SCHEMATIC DIAGRAM OF SYSTEM FOR PARAMETER OPTIMIZATION

has been added. This consists of a tachometer on the output shaft giving a voltage proportional to shaft velocity, which is fed back into the error channel through a high-pass-filter network.

The transfer function of the feedback path is given by

$$H(p) = \underbrace{(kp)}_{\text{tacho}} \underbrace{\left(\frac{Tp}{1 + Tp}\right)}_{\text{filter}} \quad \ldots\ldots [8]$$

The transfer functions of the system may now be derived

$$\frac{\theta_o}{\theta_i} = \frac{K_1(Tp + 1)}{p(T_1p + 1)(Tp + 1) + K_2Tp^2 + K_1(Tp + 1)} \quad [9]$$

[2] "Theory of Servomechanisms," by H. M. James, et al., Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, N. Y., vol. 25, 1948, pp. 333–339.

[3] Ibid., pp. 302–303, for a detailed derivation of this expression using a frequency variable ω where $p = j\omega$.

[4] Ibid., chapter 6, for a general discussion of average power expressed as an integral of a spectral density function.

where $K_2 = kK_1$.

$$\frac{\epsilon}{p\theta_i} = \frac{(T_1p + 1)(Tp + 1) + K_2Tp}{p(T_1p + 1)(Tp + 1) + K_2Tp^2 + K_1(Tp + 1)} \quad [10]$$

Expressions may now be formulated for the mean-square-error density, $\epsilon(p)\epsilon(-p)$, and for the mechanical and electrical power densities, in terms of the system parameters. The mean-square error is given by

$$\overline{\epsilon^2} = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \epsilon(p)\,\epsilon(-p)\,dp \quad [11]$$

The following parameters are fixed by the basic specification at the values given:

Input signal { mean duration of intervals, $T_2 = 4$ sec; mean square of velocities in each interval, $\overline{a^2} = 0.01$ (radians per sec)^2

Fixed elements of system { mechanical time constant, $T_1 = 1$ sec; equivalent friction resistance, $R = 250\,\Omega$; armature circuit resistance, $r_a = 100\,\Omega$; speed constant, $K_s = 1000$ v/radian per sec (including gearing to output shaft)

Power source: maximum rating of motor = 150 watts

The rms velocity of input signals, $(\overline{a^2})^{1/2}$, is 0.1 radian per sec, and substituting for the given values in Equation [3] shows that for steady-state following of this velocity the motor has a mechanical output of 40 w. There are three design parameters: namely, direct gain factor K_1, subsidiary loop gain K_2, and feedback filter time constant T, in terms of which it is required to minimize the mean-square error, $\overline{\epsilon^2}$, while keeping the total power demand $(P_f + P_r)$ to the maximum rating of the motor of 150 w.

Curve 1 in Fig. 3 shows a typical attenuation-phase analysis of the open-loop transfer function θ_0/ϵ of the system in which $T = T_1$. The gain asymptotes coincide with those for the fixed elements F, except where the latter are shown dotted. The effect of the subsidiary loop is to cause a premature —12 decibels (db) per octave slope which later "breaks" back to —6 db per octave and then joins the original curve again. It is customary to allow the 6-db section to continue for at least $2\frac{1}{2}$ octaves, which means $K_2 = 15$ db or 5, and for this simple system it is sufficient for K_1 to be chosen to meet the end of the 6-db section as shown in Fig. 3, giving $K_1 = 30$ db or 25 ($= K_2^2$). When T is less than T_1, then for K_2 and K_1 unaltered the proportions of the 12-db and 6-db sections are changed, as shown by curve 2 in Fig. 3, to the detriment of the 6-db section. It is desirable to preserve the lengths of the two sections in proportion, both increasing as T is reduced, so giving a system of adequate stability but increasing band width. This may be achieved by keeping $K_1 = K_2^2$ and varying K_2 inversely with T so that $K_2 = 5/T$ as in curve 3, Fig. 3.

Thus all the design parameters may be expressed in terms of the one variable T. Mean-square error and power now may be expressed as functions of T, using the standard integral forms given in the Appendix,[5] whence

$$\text{Mean-square error} = \overline{\epsilon^2}\ (\text{mils})^2 = \frac{0.2T^7 + 7.8T^6 + 100T^5 + 525T^4 + 964T^3}{0.000125T^8 + 0.0049T^6 + 0.0566T^5 + 0.475T^3 + 2.49T + 5} \quad [12]$$

$$\text{Mechanical power} = P_f(\text{watts}) = \frac{0.5T^3 + 19.5T^2 + 124T + 200}{0.000125T^8 + 0.0049T^6 + 0.0566T^5 + 0.475T^3 + 2.49T + 5} \quad [13]$$

$$\text{Electrical power} = P_r\ (\text{watts}) = \frac{0.2T^3 + 7.8T^2 + 51T + 110 + \dfrac{124}{T}}{0.000125T^8 + 0.0049T^6 + 0.0566T^5 + 0.475T^3 + 2.49T + 5} \quad [14]$$

These equations show that for this example with a single de-

[5] Footnote 2, p. 369, but derived in a more convenient form for the present purpose.

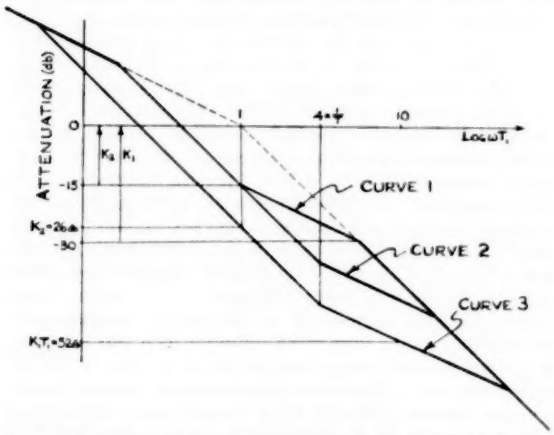


Fig. 3 Attenuation-Phase Diagram for Parameter-Optimized System

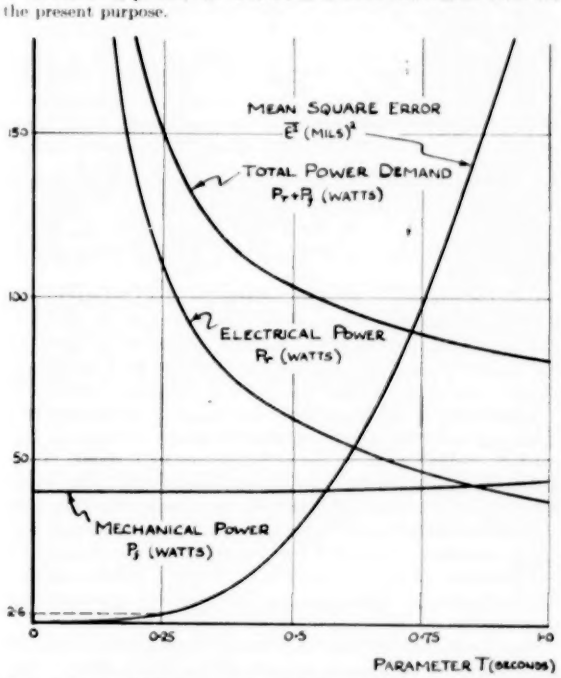


Fig. 4 Power Demand and Mean-Square Error of Parameter-Optimized System for Values of Design Parameter T

sign variable T there is a unique mean-square error $\bar{\epsilon}^2$, and power $P = P_r + P_f$, for a chosen value of the variable T. Thus there is no need to introduce the multiplier ν, as is necessary in the general case, and the graph shown in Fig. 4 of mean-square error and power against T is obtained by direct substitution for values of T in Equations [12], [13], and [14]. The mean-square error for a maximum power demand of 150 w is 2.6 (mils)^2 and occurs for $T = 0.25$ sec. The gain asymptotes for this system are shown in curve 3, Fig. 3.

The transient response of this system to a sudden change of input velocity is shown in Fig. 5. As T is reduced, the system

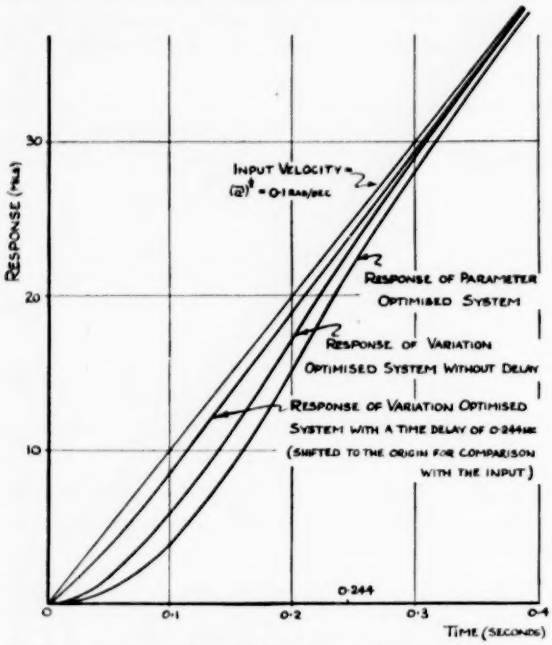


FIG. 5 TRANSIENT RESPONSES OF EXAMPLE SYSTEMS TO A SUDDEN CHANGE OF INPUT VELOCITY

band width is increased so that following becomes more precise and the mechanical power approaches that needed to follow a steady velocity equal to the rms of the input signal (that is, 40 w), the mean-square error to such a steady velocity being 0.0625 (mil)^2.

However, this is only one of the contributions to the total power demand, and the price to be paid in power demand for more accurate following of the signal is more directly shown by the electrical losses P_r which arise mainly because of the need to transfer energy to and from the load inertia if a well-damped response is to be maintained. This loss is dissipated entirely in the armature circuit and indicates clearly the need for special attention in the armature design of motors for feedback-control systems. It is of interest to note that the loss now involved is due to the feedback-controlled transactions between the energy stored in the load inertia and the power supply, and is thus only a fraction of what it otherwise would be if the system were an open-loop one with damping obtained by direct friction loss. Similar conclusions, analogous to those arrived at here, apply to power sources based on other media.

The servomechanism has been designed along standard lines and a particular solution chosen to satisfy a power limit. The question now arises of how restricting the particular choice of stabilizing method has been within the scope allowed by the power limit. This will be investigated in the next section using the variational design method.

VARIATION-OPTIMIZED SYSTEM WITH POWER LIMIT

Procedure. The method of design to be discussed here is the important contribution of this paper. It is a method based upon the calculus of variation and the use of a Lagrangian multiplier ν in terms of which a general solution for the system giving minimum mean-square error is obtained while satisfying a restraint on the power demanded. The additional elements, and their interconnections, to be added to form the system are provided by calculations in terms of the specified functions, namely, the input function, the fixed-elements transfer function, and the power limit, which are required in the same form as previously. The solution is thus no more complicated than the total complexity of the specification itself, and yet it is certain that this solution cannot be improved upon by any other linear system. This is a most interesting feature of the method since it shows that complexity in itself, beyond a certain point, cannot lead to improvement.

The general relationship between the specified functions and the equivalent cascade-system function $C(p)$ which, in conjunction with the fixed elements, gives a minimum mean-square error, is derived in the Appendix.[*] The analysis involves integration of functions of a complex variable and an application of variational calculus, but is quite straightforward. It is emphasized that the derivation of the equivalent cascade-system function $C(p)$ is only a convenient mathematical step toward the final solution of the problem; it is not in itself a solution of a practical kind. It is essential to proceed a stage further and to realize the system as a feedback system. It is desirable to allow the greatest flexibility in the form of the feedback system by considering a general feedback function $H(p)$, as shown in Fig. 6. In servomechanism

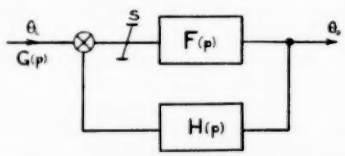


FIG. 6 SCHEMATIC DIAGRAM OF SYSTEM FOR VARIATION OPTIMIZATION

examples $H(p)$, usually but not invariably, can be realized as a direct monitoring feedback between input and output angle, together with a paralleled stabilizing feedback path.

Two basic operations are involved in solving for these optimum systems. The first is that of solving for the roots of polynominals formed by adding the constant ν (the Lagrange multiplier) to a symmetric rational function. This is a type of operation which is familiar in feedback-system calculations, since a similar operation is involved every time the effect of a feedback loop is considered (i.e., $1 + HF$). It is easiest carried out by a p-plane analysis.[6] The other operation is the one of finding the "backward" part of a function of a complex variable; i.e., the part of the function which operates on the past and hence is "backward looking." This part of the function regarded as a system function gives rise to a transient response for $t > 0$ only. When input and output are required to approximate simultaneously (without time shift), there is no difficulty with this operation which then involves taking residues at singularities in the left half p-plane only. When time delay is allowed between output and input, there is a difficulty which arises owing to the overhang in the backward part from residues at poles in the right half

[6] Also known as the root-locus method.

plane in the presence of a shift factor. However, with servomechanisms where small delays alone are advantageous, it is frequently the case that a negligible difference results if this overhang is neglected and again only the residues at singularities in the left half plane are taken into account.

Example (continued). The variation method is now applied to the example of the preceding section. The fixed elements need more precise definition in this case since the amplifier is not strictly a prescribed part of the system, and the motor alone is definitely specified giving a transfer function for fixed elements

$$F(p) = \frac{1}{K_s\, p(T_1 p + 1)}$$

where

T_1 = mechanical time constant
= 1 sec
K_s = speed constant = 1000 v/radian per sec

$$= \frac{1}{10^3\, p(p+1)} \qquad [15]$$

The input signal spectrum $G(p)$ is given from Equation [2]

$$p\,G(p) = \frac{(2\overline{a^2}\,T_3)^{1/2}}{(T_3 p + 1)}$$

where

$\overline{a^2}$ = mean-square velocity of input
= 0.01 (radian per sec)^2
T_3 = mean duration of interval
= 4 sec

whence

$$G(p) = \frac{(0.08)^{1/2}}{p(4p+1)} \qquad [16]$$

Both F and G are backward operators.

In order to evaluate the equivalent cascade system function $C(p)$ as given in Equation [33] of the Appendix, the two operations discussed in the last section must be carried out as follows:

(*a*)[7] Derivation of $F_\nu \bar{F}_\nu$

$$F_\nu \bar{F}_\nu = \nu + F\,\bar{F}$$

where ν is a Lagrangian multiplier, a positive constant. Thus

$$F_\nu \bar{F}_\nu = \nu + \frac{1}{10^6 p^2 (p+1)(p-1)} = \frac{\nu 10^6 p^2 (p^2-1)+1}{10^6 p^2 (p+1)(p-1)}$$

Let $\nu\, 10^6 = 1/r^4$, then

$$F_\nu \bar{F}_\nu \doteq \frac{\dfrac{p^4}{r^4}+1}{10^6 p^2 (p+1)(p-1)} \text{ (provided } r \gg 1)$$

$$= \left[\frac{\dfrac{p^2}{r^2} + \dfrac{\sqrt{2}p}{r} + 1}{10^3\, p(p+1)}\right]\left[\frac{\dfrac{p^2}{r^2} - \dfrac{\sqrt{2}p}{r} + 1}{10^3 p(p-1)}\right] \qquad [17]$$

(*b*) Backward part of $\bar{F}G/\bar{F}_\nu$ = sum of partial fraction terms having backward denominators

$$= (0.08)^{1/2}\,\frac{(k_1 p + 1)}{p(4p+1)} \qquad [18]$$

where

$$k_1 = \frac{\sqrt{2}/r + 1/4r^2}{1 + \sqrt{2}/4r + 1/16r^2}$$

[7] Unbarred functions are backward operators; e.g., $F(p)$. Barred functions are forward operators; e.g., $\bar{F}(-p)$.

Thus substituting in Equation [33], the equivalent cascade system function is

$$C(p) = 10^3 p(p+1)\,\frac{(k_1 p + 1) r^2}{(p^2 + \sqrt{2}pr + r^2)} \qquad [19]$$

and the over-all transfer function of the system becomes

$$CF = \frac{(k_1 p + 1) r^2}{(p^2 + \sqrt{2}pr + r^2)} \qquad [20]$$

Power assessment: The output θ_o may now be written

$$\theta_o = GCF = \frac{(0.08)^{1/2}}{p(4p+1)}\,\frac{(k\,p+1) r^2}{(p^2 + \sqrt{2}\,p\,r + r^2)} \qquad [21]$$

hence substituting for θ_o in Equation [5], δP_r may be formulated, and performing the integral of Equation [6], using a standard form from the Appendix, we have

$$\underset{\text{(watts)}}{P_r} \doteq \frac{8.46r + 22 + 21/r}{1 + 1.06/r} \qquad [22]$$

When $r = 11.4$, Equation [22] gives $P_r = 110$ w and similarly $P_f = 40$ w, giving a total power demand of 150 w.

Error assessment

$$\epsilon = \theta_i - \theta_o = G(1 - CF) = \frac{(0.08)^{1/2}}{(4p+1)}\,\frac{(p + \sqrt{2}/r - k_1)\, r^2}{(p^2 + \sqrt{2}pr + r^2)} \qquad [23]$$

whence performing the integral of Equation [11], using a standard form

$$\overline{\epsilon^2} \doteq \frac{0.08}{32\sqrt{2}\, r^3} = 1.2 \text{ (mils)}^2 \text{ for } r = 11.4$$

as compared with 2.6 (mils)^2 previously.

Substituting for r in Equation [20] gives the optimum system transfer function satisfying the power limit. The transient response of this system to a step velocity change equal to the rms velocity of the input is shown in Fig. 5, together with the response of the parameter optimized system. In this example the standard method of stabilizing with optimum parameters acquits itself well; however, it will not always compare so favorably when more complicated fixed elements are specified. In the example the question of instability does not arise since the fixed elements themselves do not lead to an unstable system. However, a substantial advantage, even in this example, is to be gained by the employment of time shift, as will be shown in the next example.

Physical realization of $H(p)$: In the physical realization of the feedback function $H(p)$ (see Fig. 6), it is desirable to allow a greater gain in the forward path than K_s, the speed constant of the motor alone. At this stage the amplifier is reintroduced so that the forward-path transfer function becomes

$$F_1(p) = \frac{K_1}{p(p+1)}$$

It is important, however, that the over-all transfer function CF remain unaltered which will be the case provided C is divided by the change of gain, namely, $K_1 K_s$, when it is associated with $F_1(p)$; thus

$$\frac{C(p)}{K_1 K_s} = \frac{1}{1 + H(p) F_1(p)}$$

whence

$$H(p) = \frac{1}{F_1(p)}\left[\frac{K_1 K_s}{C(p)} - 1\right] \qquad [24]$$

giving

$$H(p) = \frac{p^2 + \sqrt{2}\,pr + r^2}{(k_1 p + 1)r^2} - \frac{p(p+1)}{K_1}$$

Now K_1 may be chosen to give the most convenient form of $H(p)$:

When $K_1 = 4r^2$ then

$$H(p) = \underbrace{\frac{\sqrt{2}}{r}\,\frac{k_1 p^2}{(k_1 p + 1)}}_{(a)} - \underbrace{\frac{p^2}{4r^2}}_{(b)} + \underbrace{1}_{(c)} \quad \ldots\ldots [25]$$

The three components of $H(p)$ are as follows:

(a) Tachometer feedback with a high-pass filter, as used previously.

(b) Positive acceleration feedback.

(c) Monitoring loop, giving comparison between input and output position. Substituting for $r = 11.4$ gives

$$H(p) = 0.124\,\frac{0.122\,p^2}{(0.122\,p + 1)} - 0.00192\,p^2 + 1 \ldots [26]$$

The steady-state velocity-error constant $K_v = K_1 = 520$ sec^{-1} as compared with 400 sec^{-1} previously. This would give a mean square error of 0.037 (mils)^2 in following a steady-state velocity equal to the rms velocity of input signals. The subsidiary loop gain $K_2 = 64.5$, compared with 20, and the feedback network time constant is 0.122 sec as opposed to 0.25 sec. The improvement becomes possible because of the small amount of positive acceleration feedback, not previously present. The value of an accurate angular accelerometer is clear.

An Example With Time Shift. In this section it is assumed that the servomechanism of the previous example is required for an application in which a pure time delay between input and output can be allowed for. The effect of allowing a time delay in the output response of the system will be investigated and the improvement in performance resulting from this measure determined.

The input signal, fixed elements, and the function $F_v\,\bar{F}_v$ are as in the previous section. The backward operator, however, is modified by the time shift t; the overhang from the forward part of the function, resulting from time shift in Equation [33] of the Appendix, is neglected.

Backward part of $(G\bar{F}/\bar{F}_v)e^{-tp}$ = sum of partial fraction terms, modified by the exponent, having backward denominators which is given by

$$\frac{(0.08)^{1/2}\,(k_2 p + 1)}{p(4p + 1)} \quad \ldots\ldots [27]$$

where

$$k_2 = 4\left(1 - \frac{e^{t/4}}{1 + \sqrt{2}/4r + 1/16r^2}\right)$$

and is dependent upon t, the shift.

Whence the over-all system transfer function is

$$CF = \frac{(k_2 p + 1)r^2}{(p^2 + \sqrt{2}\,pr + r^2)} \quad \ldots\ldots [28]$$

for time greater than t.

Power assessment: The output θ_o is now given by

$$\theta_o = GCF = \frac{(0.08)^{1/2}\,(k_2 p + 1)r^2}{p\,(4p + 1)\,(p^2 + \sqrt{2}\,pr + r^2)}$$

Hence by Equations [5] and [6], using a standard form, we have

$$P_r = \frac{k_2^2 r^2 (r^2 + r^2/2\sqrt{2} + r) + r^2 + 4\sqrt{2}r^2 + r}{32\sqrt{2}\,r^2 + 16\,r + 2\sqrt{2}} \text{ (watts)} \quad [29]$$

which, involving k_2, also depends upon the time shift t, and so cannot yet be determined.

Error assessment: The error has to be determined with respect to the input signal delayed t seconds

$$\epsilon = (\theta_i \text{ shifted back } t) - \theta_o = G\,(e^{-tp} - CF)$$

$$= \frac{(0.08)^{1/2}}{p(4p + 1)}\left[e^{-tp} - \frac{(k_2 p + 1)r^2}{(p^2 + \sqrt{2}\,pr + r^2)}\right] \quad \ldots\ldots [30]$$

It is now necessary to perform the Integral [11], and the standard forms are not suitable, owing to the presence of the exponential shift factor. The method of residues has to be employed with attention to the contribution near the origin. The mean-square error is given by

$$\bar{\epsilon^2} \doteq 0.08\left\{t + \frac{k_2^2}{8} - \frac{\sqrt{2}\,k_2^2}{64\,r} + \frac{k_2^2}{512\,\sqrt{2}\,r^2} + k_2 - \frac{\sqrt{2}}{r}\right.$$

$$\left. - \frac{1}{32\sqrt{2}\,r^2} + e^{\frac{-tr}{\sqrt{2}}}\left[\frac{k_2}{8r^2}\cos\frac{tr}{\sqrt{2}} - \left(\frac{k_2 - \sqrt{2}}{8r^2}\right)\sin\frac{tr}{\sqrt{2}}\right]\right\}$$

$\ldots\ldots [31]$

The slope of this function at $t = 0$

$$\left.\frac{\partial \bar{\epsilon^2}}{\partial t}\right|_{t=0} \doteq -\frac{1}{16r^2}$$

signifying an initial decrease in error. However, for t large, the slope is positive. When r is regarded as a constant (which does not correspond to constant power demand) the minimum point is near

$$t \doteq \frac{2\sqrt{2}}{r} - \frac{1}{2r^2} + \frac{1}{6\sqrt{2}\,r^3}$$

and for this value of t, $k_2 = -k_1$ (of the preceding section). Thus $k_2^2 = k_1^2$ and the expression for P_r reduces to the same expression as in the previous section, which means the same value of $r = 11.4$ corresponds to a 150-w total demand. Also, since $k_2 = -k_1$, the over-all system transfer function CF is the same as in Equation [20], except that the zero lies in the right half plane

$$CF = \frac{(-k_1 p + 1)r^2}{(p^2 + 2pr + r^2)} \quad \ldots\ldots [32]$$

for time greater than t.

This convenient relationship between the two over-all transfer functions provides a simple condition for determining the most favorable time shift. The transient response of this last Function [32] to a velocity step equal to the rms of input velocities is shown in Fig. 5, shifted back to the time origin, in order to provide a comparison with the responses of the previous systems.

Substituting for k_2 in Equation [31] gives

$$\bar{\epsilon^2} \doteq 0.08\,\frac{0.215}{32\sqrt{2}\,r^2}$$

and for $r = 11.4$

$\bar{\epsilon^2} = 0.257$ (mils)^2 as compared with 1.2 (mils)^2 without shift.

Appendix

Table of Standard Integral Forms for Evaluation of Mean-Square Values

$$S_n = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\left[\frac{g_n(p)}{h_n(p)}\right]\left[\frac{g_n(-p)}{h_n(-p)}\right]dp$$

where

$$h_n(p) = a_0p^n + a_1p^{n-1} \ldots a_n$$

$$g_n(p) = d_0p^{n-1} + d_1p^{n-2} \ldots d_{n-1}$$

$$S_1 = \frac{d_0^2}{2a_0a_1}$$

$$S_2 = \frac{\dfrac{d_0^2}{a_0} + \dfrac{d_1^2}{a_2}}{2a_1}$$

$$S_3 = \frac{\dfrac{a_2}{a_0}d_0^2 + (d_1^2 - 2d_0d_2) + \dfrac{a_1}{a_3}d_2^2}{2(a_1a_2 - a_0a_3)}$$

$$S_4 = \frac{\dfrac{d_0^2}{a_0}(a_2a_4 - a_1a_4) + a_3(d_1^2 - 2d_0d_2) + a_1(d_2^2 - 2d_1d_2) + \dfrac{d_3^2}{a_4}(a_1a_2 - a_0a_3)}{2\,(a_1a_2a_3 - a_0a_3^2 - a_1^2a_4)}$$

Derivation of $C(p)$; the Equivalent Cascade System Function

The schematic diagram of the system for variation optimization is shown in Fig. 6, from which it will be seen that

$$\frac{\theta_o}{\theta_i} = \frac{F}{1 + HF}$$

and hence may be regarded as the product CF where $C = 1/(1 + HF)$. C is then the equivalent cascade-system function.

The mean-square value of signal S is uniquely related to power demand by virtue of the fixed elements $F(p)$ being known. The mean-square value of S, $\overline{S^2}$, is employed as the constraint, where $S = GC$.

The error with time shift is given by

$$= \theta_i e^{\alpha p} - \theta_o$$
$$= G(e^{\alpha p} - CF)$$

where α is time shift which is negative for a time delay.

The expression to be minimized is then

$$T = \overline{\epsilon^2} + \nu\overline{S^2} = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}[G(e^{\alpha p} - CF)\bar{G}(e^{-\alpha p} - \bar{C}\bar{F}) + \nu\, GC\,\bar{G}\bar{C}]\,dp$$

where $G = G(p)$; $\bar{G} = G(-p)\ldots$, and ν is a constant

$$= \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} G\bar{G}\,(1 - CFe^{-\alpha p} - \bar{C}\bar{F}\,e^{\alpha p} + C\bar{C}F\bar{F} + \nu C\bar{C})\,dp$$

Let δT = variation of T, ξ = a parameter independent of $C(p)$, $\eta(p)$ = variation of $C(p)$; then

$$T + \delta T = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} G\bar{G}\,[1 - (C + \xi\eta)Fe^{-\alpha p} - (\bar{C} + \xi\bar{\eta})\bar{F}\,e^{\alpha p} + (C + \xi\eta)(\bar{C} + \xi\bar{\eta})F\bar{F} + \nu(C + \xi\eta)(\bar{C} + \xi\bar{\eta})]dp$$

whence

$$\left.\frac{\partial[\delta T]}{\partial\xi}\right|_{\xi\to 0} = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\{\eta G\bar{G}[(\nu + F\bar{F})\bar{C} - Fe^{-\alpha p}] + \bar{\eta}G\bar{G}[(\nu + F\bar{F})C - \bar{F}\,e^{\alpha p}]\}\,dp$$

For a minimum, the variation represented by this integral must be zero, which it will be if each part of the integrand is analytic in one half plane. Consider the second part of the integrand which may be written, replacing $(\nu + F\bar{F})$ by $F_\nu\bar{F}_\nu$

$$\bar{\eta}\bar{G}\bar{F}_\nu\left(GF_\nu C - \frac{G\bar{F}\,e^{\alpha p}}{\bar{F}_\nu}\right)$$

This function exists entirely in the right half plane of p, provided the part in parentheses has no backward part. This is the case, and expresses the conditions for a minimum, when

$$GF_\nu C = \mathfrak{B}\left\{\frac{G\bar{F}}{\bar{F}_\nu}\,e^{\alpha p}\right\}$$

where $\mathfrak{B}$ signifies "the backward part of." Or

$$C = \frac{1}{GF_\nu}\,\mathfrak{B}\left\{\frac{G\bar{F}}{\bar{F}_\nu}\,e^{\alpha p}\right\} \quad \ldots\ldots\ldots\ldots\ldots [33]$$

where

$$\mathfrak{B}\left\{\frac{G\bar{F}e^{\alpha p}}{\bar{F}_\nu}\right\} = \frac{1}{2\pi j}\cdot\int_0^\infty e^{-pt}\,dt\int_{-j\infty}^{+j\infty}\frac{G\bar{F}}{\bar{F}_\nu}\,e^{(\alpha+t)p}\,dp$$

($\alpha = -v\tau$ for a time delay.)

Discussion

R. Oldenburger.[8] This paper is an outstanding contribution to the science of automatic control. It is gratifying to see the isoparametric problem of the calculus of variations employed in automatic-control theory. The introduction of branches of mathematics, well-known to the mathematician but new to the area of engineering involved, often, if not always, yields results significant to industry. This paper emphasizes the practicability of statistical methods in treating automatic control problems, and feedback problems in general. Statistical methods have been ably publicized by Norbert Wiener. Although, in so far as we have been able to learn, Wiener's theory on the subject has not yet been applied to industry, Dr. Westcott's paper indicates that we may expect such applications of the theory in modified form.

Servomechanisms normally are designed to run into power saturation for large signals. The reader might consider this to be a definite limitation on the use of the author's theory. This is not the case. One can employ this theory to determine the average power required by a servomechanism and design accumulators or other energy-storage means so that the maximum power available for large upsets is in excess of the average power by a safe amount. In a regulator problem the signal is normally zero;

[8] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

that is, the controller setting is fixed. As it stands, the author's theory does not apply to regulators as a whole, but may be useful for a servomechanism incorporated in the regulator, in so far as it must respond to corrective signals that arise from the disturbances. To make the paper understandable to a wider audience some points need clarification. Formula [2] as written does not occur in the book by James, et al. (author's footnote 2), to which Phillips, in James, et al., gives

$$G_x(f) = a^2 \frac{4\beta}{\beta^2 + (2\pi f)^2} \quad [34]$$

The transition to the author's Formula [2] by setting

$$p = 2\pi f j, \quad \beta = 1/T_3$$

runs as follows: Here p is the s of the ASME Dynamic Systems Committee recommendations given in paper 53—A-11, pp. 1155–1176. Evidently, the author factored

$$\frac{1}{2} G_x(f)$$

to obtain

$$\frac{2a^2\beta}{\beta^2 + (2\pi f)^2} = \frac{2\overline{a^2}\, T_3}{1 - T_3^2 p^2} = \left[\frac{\sqrt{2a^2\, T_3}}{1 + T_3 p}\right]\left[\frac{\sqrt{2a^2\, T_3}}{1 - T_3 p}\right] \quad [35]$$

He uses $^1/_2 G_x(f)$ instead of the spectral density function and performs his integrations from $-\infty$ to ∞ instead of 0 to ∞ as shown later. The author represents the function $^1/_2 G_x(f)$ by the factor

$$\frac{\sqrt{2a^2\, T_3}}{1 + T_3 p} \quad [36]$$

which he denotes by $p\theta_i(p)$.

To explain this notation, let $G_i(f)$ be the spectral density function of the input signal. We assume that we can factor $^1/_2 G_i(f)$ into

$$\frac{1}{2} G_i(f) = \theta_i(p)\theta_i(-p) \quad [37]$$

This function $\theta_i(p)$ must not be confused with the Laplace transform of the input signal. We recall that if the input to a physical system with a transfer function $Y(p)$ has the spectral density function $G_i(f)$, the spectral density function $G_x(f)$ of the output is

$$G_x(f) = |Y(p)|^2 G_i(f) \quad [38]$$

(see James, et al.). The assumption that $^1/_2 G_i(f)$ factors as in Equation [37] of this discussion appears to be a very serious restriction on $G_i(f)$. The question arises: For how general a function $G_i(f)$ is the factorization, Equation [37], valid?

We take the derivative of the input signal. The spectral density function $G_x(f)$ of this derivative is $|p|^2 G_i(f)$ since the transfer function of the derivative is p. Note that $|p|^2 = (-p)p$. Thus

$$\frac{1}{2} G_x(f) = [p\theta_i(p)]\,[-p\theta_i(-p)] \quad [39]$$

In the foregoing sense we can represent $^1/_2 G_x(f)$ by the factor $\theta_x(p)$ where

$$p\theta_x(p) = \frac{\sqrt{2\overline{a^2}\, T_3}}{1 + T_3 p} \quad [40]$$

as in the author's Formula [2].

By James, et al., if $y(t)$ is the average voltage across a unit resistance, the average power for this resistance is $\overline{y^2}$ where

$$\overline{y^2} = \operatorname*{Lim}_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y^2(t)dt \quad [41]$$

Phillips in James, et al., shows that this average power is

$$\operatorname*{Lim}_{T\to\infty} \frac{1}{T} \int_0^\infty |A_T(f)|^2 df \quad [42]$$

where

$$A_T(f) = \int_{-T}^{T} y(t) e^{-2\pi j f t}\, dt \quad [43]$$

Thus this power is

$$\int_0^\infty G(f)df \quad [44]$$

for

$$G(f) = \operatorname*{Lim}_{T\to\infty} \frac{1}{T} |A_T(f)|^2 \quad [45]$$

where $G(f)$ is the spectral density function of $y(t)$.
If

$$\operatorname*{Lim}_{T\to\infty} \frac{A_T(f)}{\sqrt{T}} \quad [46]$$

exists, and we identify the author's input signal with the $y(t)$ of James, et al., the factor $\sqrt{2}\theta_i(p)$ of $G_i(f)$ may be taken to be the limit in Equation [46]. The author's function $e_b(p)$ is a factor of the spectral density function of the voltage e_b. The spectral density function $G_b(f)$ of e_b is thus

$$2e_b(p)e_b(-p) \quad [47]$$

whence the average power P_f for the resistance R is

$$P_f = \frac{1}{R} \int_0^\infty G_b(f)df = \frac{1}{2R} \int_{-\infty}^{\infty} G_b(f)df \quad [48]$$

in view of

$$G_b(f) = G_b(-f) \quad [49]$$

Since $p = 2\pi j f$ this yields

$$P_f = \frac{1}{4\pi j} \int_{-j\infty}^{j\infty} G_b(p)dp \quad [50]$$

From this we have the author's equation preceding his Formula [3], and thus derive Equation [4].

The expression $\theta_o(p)$ is a factor of the spectral density function $G_o(f)$ of the servomechanism output θ_o where

$$\frac{1}{2} G_o(f) = \theta_o(p)\theta_o(-p) \quad [51]$$

That the author's Equation [9] holds for $\theta_i(p)$ and $\theta_o(p)$ follows easily, and the same is true of his Equation [10], where if the spectral density function of the error is $G_\epsilon(f)$, then ϵ in his equation is the $\epsilon(p)$ for which

$$\frac{1}{2} G_\epsilon(f) = \epsilon(p)\epsilon(-p) \quad [52]$$

His Equation [11] is simply

$$\overline{\epsilon^2} = \frac{1}{2} \int_{-\infty}^{\infty} G_\epsilon(f)df \quad [53]$$

To obtain his Equation [18] the rational function $\bar{F}G/\bar{F}_\nu$ is split into partial fractions. The partial fractions with poles in the left-hand plane (and at the origin) are collected into one term, given in his Equation [18], and form the "backward" part of $\bar{F}G/\bar{F}_\nu$. The rest of $\bar{F}G/\bar{F}_\nu$ is then the "forward" part with the poles in the right half plane.

In his section entitled, An Example With Time Shift, the author allows for a pure time delay between the input and the output of the servomechanisms. This is not to be confused with the delay produced by a physical component that is part of the servomechanism. Rather, this time delay is a purely mathematical one in that the output θ_o [this is the real output] is compared with the input θ_i at an instant t units of time (such as t seconds) earlier. There are thus situations, as in the case of a profiling machine, where the output does not have to follow the input instantaneously, but one is satisfied if the output is the same as the input except for a constant time delay, such as $^1/_2$ sec.

The "overhang" referred to by the author in the paragraph preceding his Equation [27] refers to the fact that the operator $e^{\alpha p}$ shifts the values of a transfer function by which it is multiplied.

In the treatment of $C(p)$ in the Appendix, the author states that the assignment of a value to the power demand P is equivalent to assigning a value to the mean-square value $\overline{S^2}$ of the signal S. This is true if the power is supplied to electrical resistances, as in the example in Fig. 1 of the paper. What if this is not the case? The author's theory assumes that the power demand is related to the input signal by a transfer function. Although this is true in many problems, this assumption appears to the writer to be a definite restriction on the generality of the theory.

The results of the argument involving the minimization of the integral T are correct, although the argument itself is not rigorous. The expression T is not to be minimized. Rather ϵ^2 is to be minimized subject to the restriction that S^2 is a constant. This is the isoparametric problem of the calculus of variations and ν is Lagrange's multiplier λ. Instead of replacing C by $C + \xi\eta(p)$ it is necessary to replace C by $C + \xi_1\eta_1(p) + \xi_2\eta_2(p)$ as the following argument shows:

Consider a function $f(x, y)$ which is to be minimized while a second function $g(x, y)$ is kept equal to a constant so that

$$g(x, y) = K \quad [54]$$

Assume that x is the independent variable, that is, that Equation [54] can be solved for y. Then f is a function of x. Setting df/dx equal to zero we obtain

$$f_x + f_y y_x = 0. \quad [55]$$

where the subscripts denote derivatives. Differentiating both sides of Equation [54] with respect to x we have

$$g_x + g_y y_x = 0. \quad [56]$$

Equations [55] and [56] are linear equations in the "variables" 1 and y_x, whence (assuming they are satisfied) the determinant of coefficients

$$\begin{vmatrix} f_x & f_y \\ g_x & g_y \end{vmatrix} \quad [57]$$

vanishes. We introduce f^* where

$$f^* = f + \lambda g. \quad [58]$$

Setting $\partial f^*/\partial x$ and $\partial f^*/\partial y$ equal to zero as if we were minimizing f^* we have

$$\left.\begin{aligned} f_x + \lambda g_x &= 0 \\ f_y + \lambda g_y &= 0 \end{aligned}\right\} \quad [59]$$

These equations are linear in 1 and λ, whence (assuming they are satisfied) the determinant of coefficients vanishes. This determinant is the same as Equation [57]. It follows that the existence of a solution of Equations [54] to [56] for x, y, and y_x, is equivalent to the existence of a solution x, y, and λ of Equations [54] and [59]. In solving Equations [54] and [59] *we are not minimizing* f^*. Neither is the author minimizing T. The integral T takes the place of the function f^*, and when C is replaced by $C + \xi_1\eta_1 + \xi_2\eta_2$ the integral T becomes a function of ξ_1, ξ_2, as well as the multiplier ν. Note that ϵ^2 is the f of Equation [59], that S^2 is the g of Equation [54], and that x, y, λ are replaced by ξ_1, ξ_2, and ν, respectively. Differentiating T with respect to ξ_1 and ξ_2 and setting the resulting derivatives equal to zero at $\xi_1 = \xi_2 = 0$, we obtain the vanishing of the integral which the author derives for

$$\frac{\partial[\delta T]}{\partial \xi}\bigg]_{\xi \to 0} \quad [60]$$

Clearly, the result at this point is the same as if we were minimizing T and therefore used one parameter ξ only.[9]

The author now assumes implicitly that the integrand of the last-mentioned integral is asymptotic to $1/p^2$ as $p \to \infty$. In referring to parts of the integrand he uses the word "exists" meaning "analytic." The rational function $\nu + F\bar{F}$ can be factored into $F_\nu\bar{F}_\nu$, where all of the poles and zeros of F_ν are in the left half plane. Poles and zeros on the imaginary axis are readily taken care of and will not be discussed here. The function

$$\frac{GFe^{\alpha p}}{F_\nu} \quad [61]$$

can be split into a sum

$$B + F \quad [62]$$

where B has its poles in the left-hand plane and F_1 its poles in the right half plane. This follows immediately by splitting the rational function $G\bar{F}/\bar{F}_\nu$ into partial fractions. By setting GF_2C equal to B, the poles of

$$GF_2C - \frac{GFe^{\alpha p}}{F_\nu} \quad [63]$$

lie in the right half plane as desired. Using the complex inversion integral

$$\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{G\bar{F}}{\bar{F}_\nu} e^{(\alpha+t)p}\, dp \quad [64]$$

the author obtains a function $f(t)$ equal to this integral, that can be written as

$$f(t) = f_1(t) + f_2(t) \quad [65]$$

(here t is not to be confused with his delay time), where

$$\left.\begin{aligned} f_1(t) &= f(t), \quad t \geq 0 \\ f_1(t) &= 0, \quad t < 0 \end{aligned}\right\} \quad [66]$$

and

$$\left.\begin{aligned} f_2(t) &= f(t), \quad t < 0 \\ f_2(t) &= 0, \quad t \geq 0 \end{aligned}\right\} \quad [67]$$

He now takes the Laplace transform of $f(t)$, which is the same as the Laplace transform of $f_1(t)$. It follows[10] from complex varia-

[9] "Calculus of Variations," by Robert Weinstock, McGraw-Hill Book Company, Inc., New York, N. Y., 1952, pp. 48–50.

[10] "Modern Operational Methods in Engineering," by R. V. Churchill, McGraw-Hill Book Company, Inc., New York, N. Y., 1944.

ble theory that this transform is analytic in the right half plane, and may be taken to be the backward part of the Function [62].

FRANCIS F. LEE.[11] The author should be complimented on his recognition that power limitation, among other things, in a feedback-control system, is a fundamental limitation, and that complexity in the controller design, beyond a certain point, cannot lead to improvement of the performance.

The author considers the servomotor as the power source. In reality, the servomotor is merely a modulator, or a valve, which controls the energy flow between the prime mover and the load. The maximum instantaneous rate of energy that can flow through the motor depends on saturation levels of various kinds, such as the maximum generator emf in a Ward-Leonard system, the setting of the relief-valve pressure in a hydraulic system of the variable-displacement pump type, or the maximum valve displacement in a valve-controlled hydraulic system.

The maximum power rating of an electric motor is the maximum mechanical power it is capable of delivering to the load for a specified temperature rise. Depending on the type of duty, there is the continuous-duty rating and the intermittent-duty rating. Above these ratings there is the peak power the motor can develop, bounded only by the saturation limits.

While the part of the average power caused by armature resistance, as considered by the author, contributes to the heating of the electric motor, and consequently is related to the rating of the motor by its efficiency, the relation between the average power and the peak power is indeterminate. In a hydraulic system, the part corresponding to the armature loss is the leakage loss. Since the hydraulic fluid is an effective coolant, the leakage loss has little to do with the power rating of the hydraulic system.

In driving a normal inertia load, little average power is needed, but the instantaneous power involved can be of considerable magnitude. In a sense we may say that the load possesses a small power factor. Thus a servomotor with either a maximum power rating or a peak power capacity equal to the computed average power demand is necessarily very much underpowered.

Although the author's interpretation of power rating of a motor is misleading and will give erroneous results, the method of minimization with constraints is valid, provided the proper constraints, such as the saturation limits, are used. It should be pointed out that G. C. Newton treated this problem in great detail in his doctorate thesis at the Massachusetts Institute of Technology in 1950. The essentials of his work have been published.[12]

W. H. SURBER, JR.[13] This paper is an interesting and well-presented example of an attempt to include the inherent nonlinearity of any actual feedback-control device in a systematic design procedure for the compensating elements. A similar approach to this problem was made by G. C. Newton.[14] The results obtained were quite analogous to those presented here, except that Dr. Newton applied one additional constraint.

The general approach used, that of controlling the probability of saturation by limiting the rms signal levels at the inputs to all saturating elements, is not the only one which may be made, however. An entirely different method of attacking this problem follows. It can be shown that what we might call the "absolute optimum" performance of a power-limited system occurs when the output device (or at least one saturating element in the system) is driven at its saturation level for the entire duration of the response to the signal; i.e., until the error is corrected to zero. If the saturation signals are switched at the correct times, then the response obtainable in this way will minimize all of the error functions used as criteria of performance, and the system operation cannot be improved upon for any other manipulation of the control signal.

Thus instead of trying to minimize the probability of saturation, a control network might be designed to operate the system in a state of saturation all of the time for a given class of signals. Where the over-all size and weight of the system are important, as in aircraft applications, the maximum utilization of the equipment becomes quite necessary and operation at saturated levels for a large percentage of the response time might be desirable.

The design procedure requires the consideration of a special class of signals, such as a randomly occurring sequence of step-amplitude or step-velocity changes. The operation, briefly, is as follows: (*a*) For a step-amplitude signal, the error actuates the control system to drive the motor into saturation; (*b*) as the error decreases and the error rate becomes larger, the manipulated variable at the input to the fixed elements is switched abruptly from one saturation level to the other so that the system reaches zero error with zero stored internal energy. For complex systems several such switching points would be required.

A disadvantage of this method of design is that it says nothing about the system stability. In fact, such a system probably would have a sustained small-amplitude oscillation in the vicinity of the error null. To eliminate this, a transition may be made to an essentially linear control system for small signals. The over-all control network required is inherently nonlinear, but it may be designed in quite a systematic manner. A linear control network, designed using conventional methods and subject to a saturating signal, would not bring the system out of saturation at the desired point in its phase space.

While this method of operation may be satisfactory for positional servomechanisms, it is probably not at all appropriate for industrial process-control systems. The stringent limitations of size and weight that make maximum apparatus utilization important do not apply to most such process controls and the required speed of response is not normally so close to the ultimate performance capabilities of the control elements. The design methods of Newton and Westcott would seem to be very satisfactory where the output-power device may be made sufficiently large so that it is feasible to operate the system in an essentially linear region.

J. M. L. JANSSEN.[15] All treatments on optimum design of control systems are necessarily incomplete as long as limitations resulting from the presence of noise or from the restricted power available are not taken into account. The noise problem has already had due consideration in the literature; the aspect of power limitation now also gets the attention it deserves. Generally speaking, one could say that power limitations are important for large signal magnitudes and that noise limitations are important when signals are weak. However, there may be cases in which both aspects are of interest and the writer wonders whether the solution of the combined problem is in principle available now, or whether the combination creates new problems.

Also, the writer wonders whether this paper, which apparently

[11] Research Engineer, Servomechanisms Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

[12] "Compensation of Feedback-Control Systems Subject to Saturation," by G. C. Newton, Jr., *Journal of The Franklin Institute*, vol. 254, 1952, part 1, pp. 281–296; part 2, pp. 391–413.

[13] Associate Professor of Electrical Engineering, Princeton University, Princeton, N. J.

[14] "Statistical Filter Theory for Feedback Systems Subject to Saturation," by G. C. Newton, Jr., presented at the IRE 1951 National Convention.

[15] Research Engineer, Royal-Dutch/Shell-Laboratory, Delft, Holland.

has been written from the servomechanism point of view, can be applied to process-control systems, since the mechanism of power limitation there is basically different from that in servos. In temperature control, for instance, there is a constant heat supply regardless of the magnitude of the control signals. Furthermore, this heat supply is modulated by the control signals. Perhaps the author could give his thoughts as to the applicability of his paper to this kind of problem.

The approach in Westcott's paper to the synthesis of optimum feedback systems is entirely different from the conventional approach. However, on the third page of the paper, two of the three parameters in the example given are eliminated by methods that are based on the conventional methods of synthesis, in which power limitations are ignored. Will the author explain this part of his paper in somewhat greater detail?

R. C. Booton, Jr.[16] This paper is a clear presentation of what might be called the rms constraint approach to system optimization. The writer would call the author's attention to almost identical earlier work of G. C. Newton, Jr.[12, 14] Since some of the assumptions that are made in applying this approach are not discussed, the writer would appreciate the author's comment on the following general question:

Since consideration of physical limitations of servocomponents requires dealing with nonlinear systems, the relation between linear analysis and the actual nonlinear system must be considered. The relation between the value of the rms constraint and the value of the actual physical constraint is not discussed. In particular, it is not clear whether the rms constraint should be set so low that the system performance certainly will be almost linear or whether this constraint should be approximately equal to the maximum value of the physical variable. In the latter case, the assumption must be made that a quasi-linearization procedure is applicable.

It is unfortunate that the scarcity of specific references might leave the reader in question as to what portions of the paper are original contributions of the author.

Author's Closure

The author is very grateful to Dr. Oldenburger for his clarification of some points that were obscured by the terse style of the paper. As author, it is frequently difficult to know which points have become obscure by curtailing the analysis and which remain self-evident. Dr. Oldenburger is concerned by the restriction imposed on the spectral density function $G_i(f)$ by the requirement that it shall factorize into a symmetrical form. The author does not think this is at all a serious restriction—certainly not so serious as the requirement that a spectral density exist for the input. This latter requirement is more serious than has sometimes been acknowledged.[3] It was almost inevitable that some confusion would arise from my introduction of time delay between input and output, because of the usual association of the effect as applying to physical elements in the open-loop transfer function. Dr. Oldenburger makes the distinction quite clear.

Professor Surber introduces an interesting idea which seems to offer the best approach to the difficult problem of taking saturation effects into account. The rms constraint method of course cannot do this, as Professor Surber points out, and one should not look to this method to do so in a general way. The class C type of operation is certainly the most efficient method of operation from an energy point of view and we look forward to hearing more of Professor Surber's useful approach to efficient operations.

In reply to Dr. Janssen, it may be said that systems involving both noise and rms limits are merely a combination of the separate problems, which does not create new difficulties. There is no reason at all why the constraint should be an absolute rms magnitude; it could equally well be expressed as the rms value of a fluctuation about an arbitrary constant value. This enables, for example, the heat supply to a furnace to be constrained within a restricted rms fluctuation about a fixed value. The elimination of two parameters by an argument based on classical servo analysis in the example on parameter optimized systems was merely done to simplify the problem. Since the form of the system is chosen on practical grounds at the outset it does not alter the situation in principle if arbitrary relations between parameters are also chosen on some reasonable basis.

Messrs. Lee and Booton ably represent M.I.T. Mr. Lee gives a detailed discussion of what the author referred to as "power demand" in a motor—the use of this description was intended to cover all power dissipation involved in energy transactions between servomotor and source. He points out that the situation is different in the case of a hydraulic system, which is agreed but it is still possible to frame the constraint in such a way as to include leakage loss if desired. Attention is drawn to the work, on feedback systems subject to saturation, of G. C. Newton who makes a rather more broad application of the Lagrangian method than the author but also it is surprising to find in view of his subject constraint rms values only. Messrs. Lee and Booton are persuaded by this work that the rms constraint method is suited for dealing with saturation effects. This one finds surprising for it seems that the restrictions required on the signals to make this possible are highly artificial as compared with the practical situation.

[16] Research Engineer, Dynamic Analysis and Control Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

A Uniform Approach to the Optimum Adjustment of Control Loops

By R. C. OLDENBOURG[1] and HANS SARTORIUS,[2] MUNICH, GERMANY

The paper describes the common aspects of the problems encountered in achieving optimal adjustment in all fields of automatic control. With the aid of a linear and a quadratic measure of error the general principles are obtained by considering the output behavior that follows a normalized disturbance of the equilibrium. After examining the advantages and disadvantages of these methods as well as the limits of their applicability, a method to be termed "practical optimum" is suggested, which combines general applicability with simplicity of practical handling.

Statement of the Problem

THE term "optimum adjustment" signifies the problem of setting the adjustable parameters of a control loop so that the control action resulting after a disturbance will take place in the best possible manner. Of course, no such optimum adjustment will be universally applicable, because it is always based upon the criterion used to define optimum control action. The choice of such a criterion is rather subjective and will depend upon the special requirements of each application.

The primary object of this paper is to describe the underlying principles upon which optimum adjustments are based. Of the many possible solutions, those most important in actual practice will be demonstrated, using the so-called linear and quadratic optimizing conditions. Their advantages and disadvantages, together with the limits of their applicability, will be examined. Finally, a method to be called the "practical optimum" will be proposed. This method combines the general applicability of the quadratic optimum with the ease of handling of the linear optimum. The general derivation of the practical-optimum method, together with the detailed calculation of an example will, it is hoped, make it easy to apply this method in actual practice.

One particular aim of this paper is to present the common aspects of the problems and solutions pertinent to all fields of automatic control, i.e., both process control and servomechanisms.

Basic Assumptions and Definitions

The Control Loop. The aim of process control is to maintain some physical variable of an installation at a constant value, or to vary it according to a prescribed pattern. Changes in the set point are rare and, if they do occur, take place slowly compared to the speed of response of the control loop. Therefore, for the majority of analyses, the set point can be assumed to be constant. The process differs from the rest of the control loop in that its parameters are predetermined and fixed, and must be accepted as such when adjusting other parameters of the loop.

Servomechanisms amplify or transmit physical quantities. The input quantity to be transmitted corresponds to the set point in process control. In this case, however, the assumption of a constant set point is no longer valid as the input usually is subject to rapid variations. With servomechanisms, ordinarily it is not necessary to divide the control loop into "process" and "controller."

Both systems are governed basically by the same laws and can be represented by a block diagram such as that shown in Fig. 1. The individual elements can be arranged in various ways,

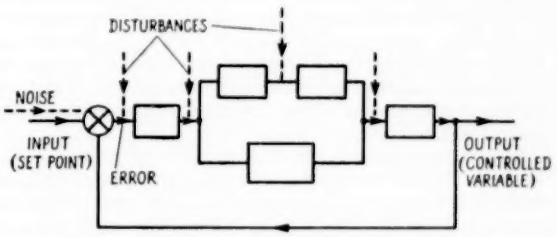


Fig. 1 Block Diagram of Generalized Automatic Control System

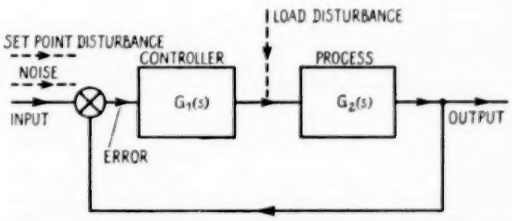


Fig. 2 Simplified Block Diagram of Control System

and, under given circumstances, also could act in a different direction. However, it is essential that a closed loop be formed with a unique continuous path of operation. Sometimes it even becomes impossible to differentiate exactly between a process-control system and a servomechanism.

As a primary consideration, the exact composition of the elements is of no consequence for the adjustment of the control loop. It is important only to differentiate between those elements or quantities which can be varied from those which cannot be modified. This consideration, together with the effect of process load disturbances which will be studied later, is the reason for dividing the control loop into the controller having the transfer function $G_1(s)$, and the process with the transfer function $G_2(s)$. Hence, we will use the simplified block diagram in Fig. 2 as a basis for the following discussion.

Degrees of Freedom. If one assumes a system consisting of lumped parameters, then the behavior of any linear control loop can be described by an ordinary differential equation of the nth order. The control problem is then also a problem of the nth order. From the physical but not the mathematical point of view, such a system will have n degrees of freedom which may be represented by the n coefficients of the characteristic equation, or

[1] Managing Director, R. Oldenbourg, Inc. (publisher).

[2] Editor, "*Regelungstechnik*."

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 11, 1953. Paper No. 53—A-18.

by the n roots of the characteristic equation. However, in any actual control system, not all of these degrees of freedom will represent parameters that can be changed at will, since some of them will be predetermined and fixed by the constructional features of the system. Therefore in control problems it becomes useful to divide the degrees of freedom into free and fixed parameters. The number of free parameters will then indicate the number of the degrees of freedom in the system representing parameters which can be chosen and modified at will, and which are therefore available for purposes of adjustment. The number of fixed parameters, on the other hand, indicates the number of degrees of freedom representing parameters which are already determined by constructional features or by the process. The sum of the number of free parameters and the number of fixed parameters must always equal the total number of the degrees of freedom.

Disturbances. In this discussion we are concerned only with stable control loops. In the absence of any forces tending to disturb the equilibrium, such systems will remain permanently at rest. Any adjustment procedure must be based on some desired shape of the recovery curve, which, for any given control loop, must in turn be based on a specified form of disturbing function. Disturbances may be classified as to location and type, that is, according to their point of action in the control loop, and according to their time variation.

Disturbances can enter at any point in the control loop, Fig. 1. In the field of process control we usually limit ourselves to the study of load disturbances, and in the field of servomechanisms to set-point disturbances, Fig. 2. These differences, of course, show up in the closed-loop transfer function G_{CL}. For pure load disturbances

$$G_{CL} = \frac{G_2}{1 + G_1 G_2} \quad [1a]$$

For pure set-point disturbances

$$G_{CL} = \frac{G_1 G_2}{1 + G_1 G_2} \quad [1b]$$

The consideration of other disturbances does not involve any basic difficulties but requires that the block diagram in Fig. 2 be broken down further into additional elements.

Classifying disturbances according to their time behavior, we differentiate between isolated and continuous disturbances. Isolated disturbances are those which occur so infrequently that their effects may be regarded as having disappeared by the time the next disturbance arises. It is therefore sufficient to make the adjustments on the basis of the effect produced by a single disturbance. Such isolated disturbances may differ widely in their time behavior. However, as one necessary characteristic, they must all have a constant (and possibly a vanishing) final value for $t \to \infty$. In calculations, isolated disturbances are usually represented by their most drastic forms, namely, step functions or impulse functions. This produces not only a mathematical simplification but also provides the most critical measure for the evaluation of the recovery curve.

In the case of continuous disturbances, the magnitude of the disturbances changes so frequently that the control loop never returns to rest. Special requirements will then apply for the choice of proper adjustment settings. Continuous disturbances may be periodic but also may be any other type of time function. Prevalent in the field of servomechanisms, a particular type of parasitic random disturbance (noise) is also of importance. Noise, however, also can play a role in process control, for example, in the form of vibration.

Quality of Control Criteria and Optimum Adjustment. The analytic adjustment for optimum results is carried out in two steps: (*a*) by the definition of a suitable quality criterion and the calculation of the corresponding measure of error. This is followed (*b*) by the determination of those values for the adjustable parameters which will produce the optimum value of the quality criterion, i.e., the minimum value of the measure of error.

It will hardly be possible to establish a method for finding optimum settings which will be equally suited for all control problems, because the requirements and demands often will vary widely. The minimum deviation, shortest recovery time, absence of oscillations, and minimum average error are some of the requirements upon which varying amounts of emphasis may be placed, depending upon the application.

In spite of the many possible criteria, only two definitions of the measure of error have generally established themselves in practice. They are the linear and the quadratic measures of error which will be considered in greater detail in the following discussion.

The measure of error lends itself to a direct minimization only if it is a function of as many variables as the number of the degrees of freedom. The quadratic measure of error fulfills this requirement. The actual optimization is then carried out by finding those physically realizable values for the "free parameters" which will make the measure of error a minimum. These are most readily found by means of the requirement that the partial derivatives of the measure of error with respect to the free parameters must vanish. Sometimes it is more convenient to find the partial derivatives not only with respect to the free parameters but with respect to all the degrees of freedom. The fixed parameters, of course, must then be considered by setting up a corresponding number of additional conditions. In any case, the optimum conditions are obtained in the form of a set of equations, equal in number to the number of free parameters of the control system. From these equations the desired values for the free parameters finally can be determined.

The measure of error can be an explicit function of only some of the system parameters, or even, as the linear measure of error, a function of only one of them. In such a case the interrelation of the various parameters must be taken into consideration. For this purpose we have available first of all the characteristic equation of the system, although this equation specifies the interrelation only in a very indeterminate form. Hence it becomes necessary to introduce additional limiting conditions. The simplest way to accomplish this is to confine the roots of the characteristic equation to certain regions of the left half complex plane. However, we must not expect to be able to obtain an exactly defined optimum recovery curve for every case, because, although the roots of the characteristic equation are specifically interrelated by this method, they are not yet bound by the initial conditions of the process under examination. The definition of the practical optimum, to be described later, takes this fact into account.

Control Loops With Isolated Disturbances

Systems With Finite and With Zero Steady-State Output. In order to achieve a uniform approach toward the effect of isolated disturbances, we must (aside from the order of the system) make certain basic assumptions about the recovery curve of the controlled variable c. We denote it in its most general form as $c(t)$ and specify at first only that the output transient $c(t)$ must vanish as $t \to \infty$. [Translator's note: $c(t)$ is used in the sense of a deviation from its initial value before the disturbance was applied.] This condition, in turn, automatically specifies what type of disturbing functions can be employed, depending on whether the closed control loop has a finite or a zero steady-state output.

It is well known that the closed-loop transfer function of a control system with zero steady-state output becomes zero for zero frequency, assuming a load disturbance. However, for a system with finite steady-state output (always the case for a set-point disturbance), $G_{CL}(0)$ assumes a finite positive value. If we now transform $c(t)$ into the complex domain by means of the Laplace transformation, we obtain

$$\mathfrak{L}\,[c(t)] = C(s) = U(s)\,G_{CL}(s) \quad [2]$$

where $C(s)$ is the Laplace transform of the transient response, represented as the product of $U(s)$, the Laplace transform of the input disturbance, and $G_{CL}(s)$, the closed-loop transfer function as defined by Equation [1*a*] or [1*b*]. Since we have specified the condition $\lim_{t \to \infty} c(t) = 0$, we obtain by means of the final value theorem the relation

$$\lim_{s \to 0} s\,C(s) = 0 \quad [3]$$

Because of certain restrictions which are to be set later, the function $C(s)$ itself must possess no zero for $s = 0$ nor, for stability reasons, a pole in the origin. In order to satisfy the requirement of a uniform type of recovery curve, and in consideration of the section on Disturbances, we decide that the following disturbances are to be used. For systems with finite steady-state output, the disturbance shall be unity in the complex domain (a unit impulse in the time domain). For systems with zero steady-state output, the disturbance shall be $1/s$ in the complex domain (a unit step in the time domain).

Linear Optimum Criterion. For the linear measure of error L, we use the so-called "linear control area," that is, the area enclosed by the recovery curve $c(t)$ and the time axis

$$L = \int_0^\infty c(t)\,dt \quad [4]$$

Under the conditions stated in the previous section, this integral will converge for all practical applications. It is evident that this measure of error is meaningful as an indication of the recovery time and of the magnitude of the deviation, but only as long as $c(t)$ assumes essentially only positive values. This requirement can be fulfilled by specifying certain limitations for the roots of the characteristic equation. Such limitations, for example, might state that the roots must be confined to certain regions of the complex plane, or that only negative real roots are permitted. The latter stipulation guarantees an absolutely nonoscillatory, though somewhat sluggish, recovery curve.

By means of the Laplace transformation, a general solution for the linear control area can be accomplished simply even without a knowledge of the specific recovery curve. Thus

$$L = \int_0^\infty c(t)\,dt = C(s)\Big|_\infty^0 \quad [5]$$

As the transfer function of every physical system must tend toward zero for $s \to +\infty$, the contribution of the lower limit of Equation [5] will always be zero, and we obtain

$$L = C(s)\Big|^0 = \lim_{s \to 0} U(s)\,G_{CL}(s) \quad [6]$$

According to this, the linear measure of error is therefore independent of frequency-variant terms of the transfer function, and thereby it also must be independent of the transient terms of the recovery curve.

By means of the measure of error thus defined, the adjustment of the control loop is defined as optimum if L assumes a minimum value. For the derivation of the minimization it is necessary, as already mentioned, to set up a number of additional conditions. The derivation itself can be accomplished in a number of different ways, which are of no particular interest here. If, for example, we set up the condition that all roots of the characteristic equation have real negative values, then the condition for a linear optimum is that, for a system of k degrees of freedom, k roots of the characteristic equation must be equal to each other. In order to calculate the corresponding parameters of the control loop, it is more expedient to begin with the mathematical conception. This specifies that, in order to satisfy the linear optimum, the k first partial derivatives with respect to s of the characteristic equation must be satisfied, in addition to the characteristic equation itself.

Because of these general and especially simple conditions, the linear optimum is of great importance, at least as a first approximation, in spite of the reservations which have been stated. The example calculated later gives more detailed information.

Quadratic Optimum Criterion. In all cases where we let the optimum recovery curve $c(t)$ assume both positive and negative values, the linear control area loses its meaning as a measure of error. This is true unless one specifies sufficient additional conditions to make sure that the recovery curve has certain characteristics. In order to be free from these limitations, it is best to select a criterion which is derived from the square of the output, namely, $c^2(t)$. The advantage that a measure of error can now be defined for the recovery curve without any restricting conditions is partially offset by the fact that a direct calculation is difficult even in simple cases.

As the quadratic measure of error, we define the so-called "quadratic control area" Q as the area enclosed by the square of the output $c^2(t)$ and the time axis. Thus

$$Q = \int_0^\infty c^2(t)\,dt \quad [7]$$

To calculate Q, it is expedient to transform the time integral into a frequency integral by means of the Parseval formula. Thus

$$\int_0^\infty c^2(t)\,dt = \frac{1}{\pi}\int_0^\infty |C(j\omega)|^2\,d\omega \quad [8]$$

where $C(j\omega)$ represents the Fourier transform of $c(t)$

$$C(j\omega) = \int_{-\infty}^{\infty} c(t)\,e^{-j\omega t}\,dt \quad [9]$$

$|C(j\omega)|$ can be visualized as the frequency spectrum of the output $c(t)$. It is evident that the quadratic control area (in contrast to the linear control area) cannot be independent of the time distribution of the recovery curve. If, however, the integrand of the right side of Equation [8] is split into the two components derived from the disturbing function and the closed-loop transfer function, respectively, then we obtain

$$Q = \frac{1}{\pi}\int_0^\infty |C(j\omega)|^2\,d\omega = \frac{1}{\pi}\int_0^\infty |U(j\omega)|^2 |G_{CL}(j\omega)|^2\,d\omega \quad [10]$$

This shows that, here too, we must assume an isolated disturbance. This is because the integral cannot converge unless both $|U(j\omega)|^2$ and $|G_{CL}(j\omega)|^2$ (or at least their product) are always finite and tend to zero sufficiently fast as ω increases. This is always the case for $|G_{CL}(j\omega)|^2$ for all physically realizable systems, but cannot be the case for $|U(j\omega)|^2$ if the disturbance changes permanently.

For all control loops which consist only of lumped parameters and where the disturbing function is chosen properly (for example, as indicated in the first part of this section) it becomes possible to represent the Fourier transform $C(j\omega)$ of Equation

[9] as a proper rational fraction in terms of $j\omega$

$$C(j\omega) = \frac{b_m(j\omega)^m + \ldots + b_1(j\omega) + b_0}{a_n(j\omega)^n + a_{n-1}(j\omega)^{n-1} + \ldots + a_1(j\omega) + a_0} \quad [11]$$

where $n \geq m + 1$.

Among other conditions which this function is subject to[3] we merely mention that its poles must all lie in the upper half of the ω-plane, exclusive of its axis, and that for this reason the constant term a_0 in the denominator must not vanish. Under this assumption, the integrand $|C(j\omega)|^2$ of Equation [10] also becomes a proper rational fraction of identical structure as $C(j\omega)$ itself, although now with the real argument ω^2

$$|C(j\omega)|^2 = J(\omega^2) = \frac{B_m(\omega^2)^m + \ldots + B_1(\omega^2) + B_0}{A_n(\omega^2)^n + A_{n-1}(\omega^2)^{n-1} + \ldots + A_1(\omega^2) + A_0} \quad [12]$$

The coefficients A_ν and B_μ, as can easily be shown, are related to those of $C(j\omega)$ in the following manner

$$\left.\begin{aligned} A_\nu &= a_\nu^2 - 2(a_{\nu-1})(a_{\nu+1}) + 2(a_{\nu-2})(a_{\nu+2}) - \ldots + \ldots \\ B_\mu &= b_\mu^2 - 2(b_{\mu-1})(b_{\mu+1}) + 2(b_{\mu-2})(b_{\mu+2}) - \ldots + \ldots \end{aligned}\right\} \quad [13]$$

where, on the right side of each equation, all terms disappear whose index coefficients become < 0, or $> n$ or $> m$, respectively.

The quadratic measure of error now becomes

$$Q = \frac{1}{\pi}\int_0^\infty J(\omega^2)d\omega \quad [14]$$

Its evaluation may be carried out in various ways. The Appendix contains a number of details concerning this problem. It turns out that Q is a function of all the coefficients A_ν and B_μ, or that, by means of Equations [13], it can be expressed as a function of all the coefficients a_ν and b_μ. The quadratic measure of error therefore includes all the roots of the characteristic equation and all the initial conditions. Therefore it includes all the degrees of freedom of the system under investigation. Its minimization therefore can be carried out directly, although this raises considerable algebraic difficulties even for relatively simple cases. The example given later illustrates this further. Owing to these difficulties, the derivation of optimum adjustments in general terms promises little success if the mathematical effort is to be kept within reasonable limits. In practice, an appreciable amount of calculation, therefore, must be carried out for each individual case.

Control Loops With Continuous Disturbances

Earlier in the paper we defined as an essential characteristic of a continuous disturbance that the control loop to which it is applied can never come to rest. Such disturbances may be any arbitrary time function.

Periodic Disturbing Function. The simplest general form of a continuous disturbance is a periodic function that can be represented by a Fourier series. If one applies such a function to the input of a control loop, then, after a sufficient length of time has elapsed, the output $c(t)$ also will become a periodic function. The output will contain all frequencies that are included in the input, the respective output amplitudes depending upon the transfer function of the system. Under these conditions, it is evident that the criteria $L = \int_0^\infty c(t)dt$ and $Q = \int_0^\infty c^2(t)dt$ which were defined previously must fail in this case, since they become either indeterminate or basically cannot assume a finite value.

Therefore it becomes reasonable to choose the average time value of the output, instead of the area criterion. In order to derive significant relations, we can represent the disturbing function by means of the simplest type of Fourier series

$$u(t) = D_0 + D_1 \sin(\omega_1 t)$$

The output will then assume the form

$$c(t) = D_0|G_{CL}(0)| + D_1|G_{CL}(j\omega_1)|\sin(\omega_1 t)$$

where the phase shift is not considered. This procedure is permissible in connection with the mathematical treatment which is to follow without limiting the general applicability. Under this assumption, the linear mean value c_m measured over a full period $T = 2\pi/\omega_1$ becomes

$$c_m = \frac{1}{T}\int_{-T/2}^{+T/2} c(t)dt = D_0|G_{CL}(0)| \quad [15]$$

and is therefore basically independent of the periodic components of the disturbing function. Of interest therefore is only the linear mean output corresponding to the constant term in the disturbing function, in other words, the so-called regulation factor $\eta = G_{CL}(0)$. This factor gives the amount of attenuation produced by the system for a constant steady disturbance. For control loops with zero steady-state output, this regulation factor also will become zero; that is, the effect of a steady constant disturbance will be eliminated completely by the control action. In order to be able to include these important control loops in our discussion, we picture the disturbing function to be applied to the input of the control system through a transmission system containing a simple integrating element, having the frequency response $1/j\omega$. In this way the behavior of the system is reduced to the normal form with steady-state output.

Analogously we can calculate the quadratic-mean value $\bar{c}_m^2$

$$\bar{c}_m^2 = \frac{1}{T}\int_{-T/2}^{+T/2} c^2(t)dt = D_0^2|G_{CL}(0)|^2 + D_1^2\frac{|G_{CL}(j\omega_1)|^2}{2} \quad [16]$$

In this case the oscillation superimposed upon the constant term of the disturbing function also contributes toward the quadratic-mean output. In the most general case, each term of the Fourier series for $u(t)$ makes some contribution. In this form an analogy with the representation of the quadratic measure of error given by Equation [10] is, it is true, not immediately apparent. However, the fact that for a periodic disturbing function $u(t)$, the output $c(t)$ also will become a function of the individual discrete frequencies ($\omega = 0, \omega_1, 2\omega_1, \ldots$), gives a clear indication that the same also must be true, using a generally applicable criterion, when the input and output are represented by the frequency spectra $|U(j\omega)|$ and $|C(j\omega)|$, respectively. We shall mention these relations again briefly in the next section.

Arbitrary Disturbing Function. An attempt to apply the quadratic measure of error, given by Equation [10], to the case of an arbitrary continuous disturbance must fail because the integral diverges. However, here it also becomes possible to introduce the quadratic-mean output. This is found by relating the quadratic control area within any time interval to the length of the time interval, and then letting the two boundaries of the time interval approach infinity

[3] "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand, Inc., New York, N. Y., 1949.

$$c_m^2 = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} c^2(t)dt \quad \ldots\ldots\ldots\ldots [17]$$

To calculate this, we again transform the time integral into a frequency integral

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} c^2(t)dt = \frac{1}{\pi} \int_0^\infty M_c(\omega)d\omega \quad \ldots [18]$$

where the function $M_c(\omega)$ is defined as the spectral density of the output

$$M_c(\omega) = \lim_{T\to\infty} \frac{1}{T} |C(j\omega)|^2 \quad \ldots\ldots\ldots\ldots [19]$$

The spectral density M_c of the output again can be split into the spectral density of the disturbing function M_u and the modulus of the transfer function G_{CL}. Thus we find for the quadratic-mean output

$$c_m^2 = \frac{1}{\pi} \int_0^\infty M_c(\omega)d\omega = \frac{1}{\pi} \int_0^\infty M_u(\omega) |G_{CL}(j\omega)|^2 d\omega \ldots [20]$$

This representation also contains, as a special case, the periodic disturbing function. For any harmonic oscillation, for example, of frequency ω_1, the spectral density $M_u(\omega_1)$ then degenerates into a discrete spectral line at the abscissa ω_1. (Expressed mathematically, an impulse at the abscissa $\omega = \omega_1$.) The quadratic-mean output then obviously takes on the form given by Equation [16]. A special advantage of the representation by Equation [20] is that, under certain assumptions, the treatment of arbitrary disturbing functions, which can only be described statistically, becomes possible by means of the spectral density[4] $M_u(\omega)$.

A comparison of Equation [10] with Equation [20] shows that the quadratic control area and the "quadratic-mean output" are completely analogous concepts, differing only in the representation of the disturbing function. The disturbing function is represented by the frequency spectrum in the case of the isolated disturbance, and by the spectral density in the case of the continuous disturbance.

The process of finding optimum adjustments is in both cases completely the same, consisting of evaluating the frequency integral and then determining the minimum. Therefore, in the discussion which is to follow, we will always let Q stand for the quadratic measure of error, and, according to the particular problem, it will represent either the quadratic control area or the quadratic-mean output.

Practical Optimum Criterion

Background. The result of the foregoing considerations may be summarized briefly as follows:

Usable methods for the optimum adjustment may be carried out in a consistent manner both for process control and for servomechanisms, using any kind of disturbance, by means of the linear and the quadratic measure of error. If sufficient conditions are stated to insure an aperiodic recovery curve, then the use of the linear optimum permits an easy mathematical treatment, but the results obtained will satisfy only relatively moderate demands of accuracy and speed of response. On the other hand, the quadratic optimum is widely applicable and produces an especially fast and accurate control action, although the response may be oscillatory. Furthermore, the mathematical derivation makes more extensive demands, which result in considerable algebraic difficulties when a practical application is attempted. The example calculated later demonstrates these facts.

These aspects make it desirable to introduce a method for finding optimum adjustments which will, if possible, combine the advantages of the two methods described. In the following discussion it will be attempted to set forth a new approach in this direction, which aims particularly at a simple and practical method of dealing with the problem.

Restricted Frequency Spectrum. We start with the quadratic measure of error in its general form, as given by Equation [14]

$$Q = \frac{1}{\pi} \int_0^\infty J(\omega^2)d\omega$$

The difficulties of finding the minimum of Q originate primarily from the fact that the integration, which is not too simple, must be carried out. The evaluation of the conditions for minimum error, to be carried out next, proves even more difficult. However, the integration merely signifies the calculation of the area bounded by the positive ω-axis and the function $J(\omega^2)$ which is to be integrated. Therefore, if Q is to assume a minimum value, then we must insure that this area becomes as small as possible. Therefore we can try to avoid the integration and merely fix the integrand in such a manner that the area enclosed between it and the ω-axis will remain as small as possible. To do this, however, it is necessary that certain conditions exist which subject the integrand to the physical requirements of the problem.

The fact that $J(\omega^2)$ represents the squared frequency spectrum of the system output permits a clear formulation of the principal requirements necessary to insure fast control action and absolute freedom from resonance peaks. We assume the most unfavorable type of disturbance, namely, an impulse, which, as we know, contains all frequencies from 0 to ∞ with equal amplitude. Then in Equation [10], $|U(j\omega)|^2 = 1$, and $J(\omega^2)$ is solely represented by the squared modulus of the transfer function $|G_{CL}(j\omega)|^2$. In this representation, Q also may be regarded as the generalized form of the relations formulated in the preceding section.

The modulus of the frequency response indicates the transmission factor of the amplitudes for all the frequencies. Resonance peaks, therefore, can be avoided if $|G_{CL}(j\omega)|$ does not exceed its value for $\omega = 0$ by any considerable amount at other frequencies, Fig. 3. On the other hand, the speed of response will

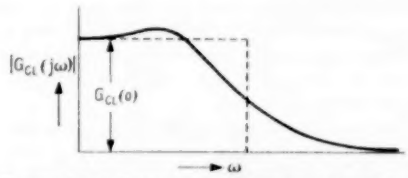


Fig. 3 Closed-Loop Frequency Spectrum of Control System

become faster as the pass band of the frequency spectrum becomes wider. These requirements will be fulfilled if the plot of $|G_{CL}(j\omega)|$ forms a rectangle with as great a length as possible. The area of this rectangle will be proportional to the quadratic measure of error Q. If $|G_{CL}(j\omega)|$ has the shape just stated, then

[4] "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

"Extrapolation, Interpolation and Smoothing of Stationary Time Series," by N. Wiener, John Wiley & Sons, Inc., New York, N. Y., 1950.

the initial value $|G_{CL}(0)|$ will become a singular measure of the area. To achieve the optimum condition, we therefore should strive to make $|G_{CL}(0)|$ as small as possible. Physically, this means that, once the character of $|G_{CL}(j\omega)|$ is prescribed, then the character of the recovery curve is determined to within narrow limits, so that the linear measure of error is again of definite significance as a criterion of the quality of control.

Special Characteristics of Spectral Function. The question now arises how this ideal case can best be approached in practice. In this connection, we first must clarify some of the special characteristics of the function $J(\omega^2)$.

We assume that $J(\omega^2)$, as given by Equation [12], is a proper rational fraction in terms of ω^2. (The procedure is also applicable when meromorphic functions occur and, in this case, even provides special advantages for the mathematical treatment.) For the sake of clarity, we substitute x for ω^2, to obtain

$$J(x) = \frac{B_m x^m + \ldots + B_1 x + B_0}{A_n x^n + A_{n-1} x^{n-1} + \ldots + A_1 x + A_0} = \frac{N(x)}{D(x)} \quad [21]$$

where $J(0) > 0$ and $n > m$.

For physical reasons, $J(x)$ must be in agreement with its mathematical definition (see Equation [10]) everywhere for $x \geqq 0$, a finite positive function which tends toward zero only for $x \to \infty$. For similar reasons, it follows that for all values of $x \geqq 0$, $N(x)$ and $D(x)$ must themselves be positive functions which increase without limit for $x \to \infty$. The behavior of the successive derivatives of J is also essential. In the following discussion the case in which the argument vanishes will be of special interest. As all odd derivatives of J with respect to ω will fundamentally vanish at $\omega = 0$, it is therefore permissible to carry out the differentiation with respect to x. We obtain

$$N(x) \equiv D(x)\, J(x)$$
$$N'(x) \equiv D'(x)\, J(x) + D(x)\, J'(x)$$

$$N^{(\nu)}(x) \equiv \binom{\nu}{0} D^{(\nu)}(x)\, J^{(0)}(x) + \binom{\nu}{1} D^{(\nu-1)}(x)\, J^{(1)}(x) + \ldots$$

$$\ldots + \binom{\nu}{\nu-1} D^{(1)}(x)\, J^{(\nu-1)}(x) + \binom{\nu}{\nu} D^{(0)}(x)\, J^{(\nu)}(x) \quad [22]$$

From this representation it becomes clear that $J(x)$ possesses derivatives of any desired order. However, there are only n nonvanishing derivatives of the denominator polynomial $D(x)$, and only m (i.e., at most $n - 1$) nonvanishing derivatives of the numerator polynomial $N(x)$. All derivatives of $J(x)$ of an order higher than $n - 1$ will then develop from $J(x)$ and from its first $n - 1$ derivatives by means of a recursion formula. Therefore it is only possible to differentiate $J(x)$ n times (counting the derivative of zero order which is $J(x)$ itself) for a given argument x. All others are necessarily fixed. The n available derivatives therefore can be considered to represent the n degrees of freedom of the control system.

Conditions for Absence of Resonance Peaks. Resonance peaks can be avoided with certainty if $J(x)$ starts with a finite value of $J(0)$ and decreases monotonically for all values of $x > 0$. Another interesting possibility is to set an upper and a lower limit for the variations of $J(x)$. In this case the mathematical treatment leads to the use of Tchebychef's polynomials and would be beyond the scope of this paper.

The foregoing requirement is fulfilled if the value of the denominator polynomial $D(x)$ exceeds the value of the numerator polynomial $N(x)$ for all values of $x > 0$. From the equation

$$J(x) = \frac{N(x)}{D(x)} = \frac{B_0}{A_0} \frac{\frac{B_m}{B_0} x^m + \ldots + \frac{B_1}{B_0} x + 1}{\frac{A_n}{A_0} x^n + \frac{A_{n-1}}{A_0} x^{n-1} + \ldots + \frac{A_1}{A_0} x + 1} \quad [23]$$

it is recognized immediately that these demands, under the strict limitation of positive coefficients, are fulfilled by the following set of conditions

$$\frac{A_\mu}{A_0} \geqq \frac{B_\mu}{B_0} \text{ or } J(0) \geqq \frac{N^{(\mu)}(0)}{D^{(\mu)}(0)}, \text{ where } \mu = 1 \ldots m \quad [24]$$

If the control loop possesses only one free parameter then, with $\mu = 1$, we obtain

$$J(0) \geqq \frac{N'(0)}{D'(0)} \quad [25]$$

Otherwise we find from the second equation of the System [22]

$$J(0) = \frac{N'(0)}{D'(0)} - \frac{D(0)}{D'(0)} J'(0) \quad [26]$$

The Condition [25] therefore can only be fulfilled if $J'(0) \leqq 0$, in which case, according to Equation [26], $J(0)$ assumes a permissible minimum value if

$$J'(0) = 0 \quad [27]$$

If there is a second free parameter then, for $\mu = 2$, we have the additional relation

$$J(0) \geqq \frac{N''(0)}{D''(0)}$$

and, by using the third equation of the System [22]

$$J(0) = \frac{N''(0)}{D''(0)} - \frac{2D'(0)}{D''(0)} J'(0) - \frac{D(0)}{D''(0)} J''(0)$$

The first optimizing condition, stated by Equation [27], and the the last two relations, of course, can be reconciled only if $J''(0) \leqq 0$, where $J(0)$ again takes on a minimum value provided

$$J''(0) = 0 \quad [28]$$

If additional free parameters exist, then the successive application of this procedure leads to the following additional conditions for optimum adjustment

$$\left.\begin{aligned} J'''(0) &= 0 \\ J^{IV}(0) &= 0 \\ &\cdot \\ &\cdot \\ &\cdot \\ J^{(\mu)}(0) &= 0 \end{aligned}\right\} \quad [29]$$

Thus we have established the conditions for the practical optimum criterion. They can be simply presented by means of the numerator and denominator coefficients of $J(x)$, as follows

$$\left.\begin{aligned} A_1 B_0 &= A_0 B_1 \\ A_2 B_0 &= A_0 B_2 \\ &\cdot \\ &\cdot \\ &\cdot \\ A_\mu B_0 &= A_0 B_\mu \end{aligned}\right\} \quad [30]$$

These signify that, at $x = 0$, as many derivatives of $J(x)$ must vanish as the control loop possesses free parameters. Furthermore, the aim of having $J(x)$ closely approach the shape of a rectangle is fulfilled, in so far as the specific problem will permit, if we assume a continuously decreasing function.

The simplicity of the optimization equations in this form should be convincing. They remain simple even if the original coefficients of Equation [13] are substituted into the system of Equations [30]. The importance of the practical optimum is especially brought out by comparing it to the conditions for the quadratic optimum which are very awkward to evaluate. Compared to the linear optimum, the practical optimum is definitely superior with respect to speed and accuracy of control. The following examples will demonstrate these facts clearly.

Practical Example

In order to be able to visualize and compare the principles which have been developed, the various optimum criteria will now be tried out on a practical example. A very simple control system has been chosen intentionally so that the basic principles will become as clear as possible, and so that they may be verified easily without an excess of calculations. The numerical results obtained from this example provide such a clear confirmation of the foregoing discussion that any further comments seem unnecessary.

The Control System. We assume a first-order process with a time constant T_s and the transfer function

$$G_2 = \frac{1}{T_s s + 1} \qquad [31]$$

and a proportional-plus-integral controller with the transfer function

$$G_1 = \frac{T_r s + 1}{T T_r s^2 + (T + \rho T_r) s} \qquad [32]$$

T represents the closing time of the valve motor, T_r the feedback time constant, and ρ the influence coefficient of the feedback. The effect of a load disturbance is to be studied. Substituting $T_s s = q$, and applying an isolated step disturbance $U(q) = 1/q$, we obtain the Laplace transformed output

$$C(q) = \frac{1}{q} \frac{G_2(q)}{1 + G_1(q)\, G_2(q)} = \frac{q + b_0}{q^3 + a_2 q^2 + a_1 q + a_0} \qquad [33]$$

where the coefficients are given by

$$\left.\begin{aligned} a_0 &= \frac{T_s}{T}\frac{T_s}{T_r} \\ a_1 &= \frac{T_s}{T} + \frac{T_s}{T_r} + \rho\frac{T_s}{T} \\ a_2 &= 1 + \frac{T_s}{T_r} + \rho\frac{T_s}{T} \\ b_0 &= \frac{T_s}{T_r} + \rho\frac{T_s}{T} = a_2 - 1 \end{aligned}\right\} \qquad [34]$$

The three degrees of freedom of the system are represented by the parameters T/T_s, T_r/T_s, and ρ. Of these, T/T_s is to be considered as a fixed parameter, because T_s represents a fixed characteristic of the process, while T stands for a characteristic of the controller which usually cannot be altered in an arbitrary fashion. T_r/T_s and ρ represent the free parameters of the control loop, whose optimum values are now to be determined by means of the various optimizing criteria. The additional requirement corresponding to the fixed parameter can easily be expressed as

$$F(a_0, a_1, a_2) = a_2 - a_1 = 1 - \frac{T_s}{T} = \text{const} \qquad [35]$$

The first task now consists of finding the sets of numbers for a_0, a_1, and a_2 which satisfy the particular optimizing criterion as well as Equation [35], with physically realizable values.

Linear Optimum Criterion. The linear measure of error $L = b_0/a_0 = (a_2 - 1)/a_0$ becomes a minimum (assuming only negative real roots) if the characteristic equation of Equation [33]

$$q^3 + a_2 q^2 + a_1 q + a_0 = 0$$

and its first two derivatives with respect to q (there being two free parameters) are satisfied. Thus

$$3q^2 + 2a_2 q + a_1 = 0$$

$$6q + 2a_2 = 0$$

Therefore the characteristic equation must possess a real triple root. From these three equations and from Equation [35] the corresponding sets of values for the coefficients a_0, a_1, and a_2 and the triple root q_0 can be determined.

Quadratic Optimum. By means of the result derived in the Appendix, we obtain for this case the following expression for the quadratic measure of error

$$Q = \frac{a_0 + (a_2 - 1)^2 a_2}{2a_0(a_1 a_2 - a_0)}$$

To minimize Q, the partial derivatives of the linear compound quantity $Q + \lambda F$ with respect to the coefficients a_0, a_1, and a_2 must be set equal to zero. λ is the Lagrange multiplier, and F is the condition specified by Equation [35]. The result consists of three equations in a_0, a_1, a_2, and λ. By eliminating λ from these, two equations

$$f_1(a_0, a_1, a_2) = 0$$
$$f_2(a_0, a_1, a_2) = 0$$

remain. These, in turn, can be combined into a fifth-degree equation in a_2 alone, which must be solved several times for its various factors. Of its five roots, three must always be eliminated as they are without physical meaning. The two remaining roots must be investigated further, after determining their corresponding values of a_0 and a_1, by means of one of the equations $f = 0$ and the condition of Equation [35]. The final selection is then carried out in accord with the most favorable control action which results. The roots of the characteristic equation which are associated with each set of values for a_0, a_1, and a_2 are finally obtained as the roots of the cubic denominator-equation of $C(q)$, as given in Equation [33].

Practical Optimum. By means of the set of Equations [13] the coefficients of the frequency spectrum become

$$\begin{aligned} A_0 &= a_0^2 & B_0 &= b_0^2 = (a_2 - 1)^2 \\ A_1 &= a_1^2 - 2a_0 a_2 & B_1 &= b_1^2 = 1 \\ A_2 &= a_2^2 - 2a_1 \\ A_3 &= a_3^2 = 1 \end{aligned}$$

The two conditions for the practical optimum, corresponding to the two free parameters, become, according to Equation [30]

$$A_1 B_0 = A_0 B_1$$
$$A_2 = 0$$

or, using the foregoing coefficients

$$(a_1^2 - 2a_0 a_2)(a_2 - 1)^2 = a_0^2$$
$$a_2^2 - 2a_1 = 0$$

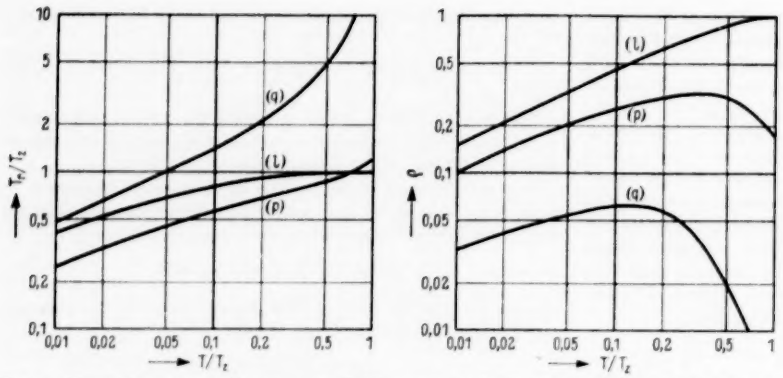


FIG. 4 FREE CONTROLLER PARAMETERS T_r/T_z AND ρ AS FUNCTIONS OF T/T_z FOR LINEAR (l), PRACTICAL (p), AND QUADRATIC (q) OPTIMA

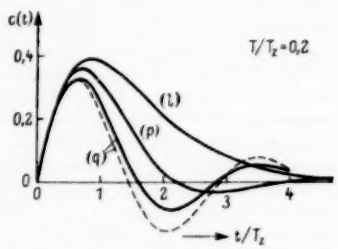


FIG. 6 TRANSIENT RESPONSE OF OUTPUT WHEN USING LINEAR (l), PRACTICAL (p), AND QUADRATIC (q) OPTIMA

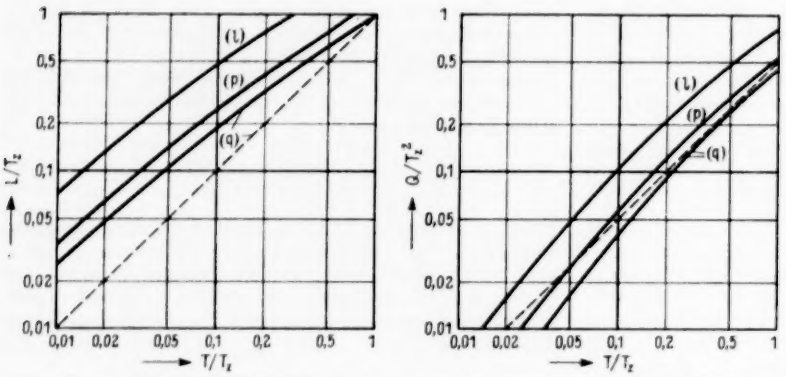


FIG. 5 LINEAR MEASURE OF ERROR L AND QUADRATIC MEASURE OF ERROR Q AS FUNCTIONS OF T/T_z FOR LINEAR (l), PRACTICAL (p), AND QUADRATIC (q) OPTIMA

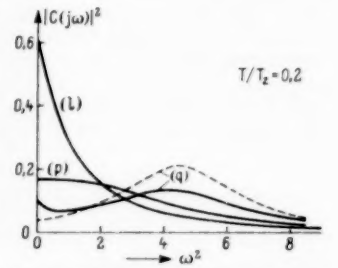


FIG. 7 FREQUENCY SPECTRUM OF OUTPUT FOR LINEAR (l), PRACTICAL (p), AND QUADRATIC (q) OPTIMA

From these relations, and the additional condition of Equation [35], the corresponding sets of values for a_0, a_1, and a_2 can be determined easily. The corresponding roots of the characteristic equation are again found from the denominator of $C(q)$ as given in Equation [33].

Numerical Calculation. For each of the three optimizing criteria, there exists a number of sets of corresponding values for the denominator coefficients of $C(q)$. Using these coefficients, and Equations [34] and [35], respectively, we can calculate easily the corresponding free parameters of the controller T_r/T_z and ρ as functions of the fixed parameter T/T_z. The results are shown in Fig. 4.

In the case of the quadratic optimum, one of the sets of values mentioned previously in this section turns out to be a special case for which ρ becomes zero, so that T_r/T_z becomes meaningless. The physical meaning of this condition is that the feedback has been removed from the controller. The results, therefore, represent the quadratic optimum for a control loop containing a process with the transfer function [31], but with a simple floating controller having the transfer function $G_1 = 1/Ts$. The corresponding numerical results are shown with a broken line in the diagrams which are to follow.

The calculation of the measures of error L and Q, obtained by means of these optimum adjustments, is carried out in the same manner. The results are plotted in Fig. 5 as a function of T/T_z.

Especially noticeable is how closely and consistently the measures of error resulting from the practical optimum approximate those resulting from the quadratic optimum. This fact is especially significant when one considers the considerably greater difficulties that must be overcome in order to calculate the quadratic optimum.

The respective recovery curves can be calculated readily by means of the corresponding roots of the characteristic equation. Fig. 6 shows the results for the specific case of $T/T_z = 0.2$.

The curves show, on the one hand, the rather sluggish control action resulting from the linear optimum and, on the other hand, the definitely oscillatory response resulting from the quadratic optimum. The practical optimum represents a happy compromise.

Finally, by means of the set of Equations [13], we find coefficients of the frequency spectrum corresponding to each set of values for a_0, a_1, and a_2

$$|C(j\omega)|^2 = J(\omega^2) = \frac{B_1\omega^2 + B_0}{A_3\omega^6 + A_2\omega^4 + A_1\omega^2 + A_0}$$

This is represented in Fig. 7 as a function of ω^2, again for the case $T/T_z = 0.2$, and for each of the three optimum criteria.

ACKNOWLEDGMENT

We are indebted to Mr. D. W. Pessen and Mr. E. F. Hochschild of the Minneapolis-Honeywell Regulator Company, Brown Instruments Division, for the translation of our paper.

REFERENCES

"The Analysis and Synthesis of Linear Servomechanisms," by A. C. Hall, M.I.T. Press, Cambridge, Mass., 1943.

"Dynamik selbsttätiger Regelungen," by R. C. Oldenbourg and H. Sartorius, R. Oldenbourg, (publisher) Munich, Germany, 1951.

"Die zweckmässige Festlegung der frei wählbaren Regelungskonstanten," by H. Sartorius, Dissertation, Technische Hochschule, Stuttgart, Germany, 1945.

"Das Optimierungsproblem in der Regelungstechnik," by H. Sartorius, *Regelungstechnik*, vol. 1, Heft 4, 1953, p. 74.

"Automatic and Manual Control," by A. Tustin, Butterworth Publications, Ltd., London, England, 1952.

"Cybernetics or Control and Communication in the Animal and the Machine," by N. Wiener, John Wiley & Sons, Inc., New York, N. Y., 1948.

Appendix

In the case where the spectral function may be broken into a proper rational fraction

$$J(\omega^2) = \frac{B_m\omega^{2m} + \ldots + B_1\omega^2 + B_0}{A_n\omega^{2n} + A_{n-1}\omega^{2n-2} + \ldots + A_1\omega^2 + A_0}$$

the calculation of the quadratic measure of error can be carried out in general form.

For this purpose, let us split $J(\omega^2)$ into a number of separate fractions, corresponding to the number of terms in the numerator polynomial

$$J(\omega^2) = \sum_{k=0}^{m} \frac{B_k\omega^{2k}}{A_n\omega^{2n} + A_{n-1}\omega^{2n-2} + \ldots + A_1\omega^2 + A_0}$$

The measure of error can thereby also be expressed as the sum of the individual parts

$$Q = \sum_{k=0}^{m} B_kQ_k$$

Each part then has the form

$$Q_k = \frac{1}{\pi}\int_0^\infty \frac{\omega^{2k}}{A_n\omega^{2n} + A_{n-1}\omega^{2n-2} + \ldots + A_1\omega^2 + A_0}\,d\omega$$

After some algebraic manipulation, the integration results in

$$Q_k = \frac{H_k^*}{2R_nR_{n-1}\ldots R_{n-(k+1)}}$$

or, written out in detail for the individual parts

$$Q_0 = \frac{H_0^*}{2R_nR_{n-1}}$$

$$Q_1 = \frac{H_1^*}{2R_nR_{n-1}R_{n-2}}$$

$$\vdots$$

$$Q_{n-1} = \frac{H_{n-1}^*}{2R_nR_{n-1}\ldots R_1R_0}$$

The symbols used have the following meaning: The characteristic equation of the system is

$$a_np^n + a_{n-1}p^{n-1} + \ldots + a_1p + a_0 = 0$$

$R_0 \ldots R_n$ are the Routh trial functions derived from the coefficients a_ν, and they are, as is well known, intimately connected with the Hurwitz determinants

$$R_\nu = \frac{H_\nu}{H_{\nu-1}}$$

The values for H_0 and R_0 are here defined as 1, and the nth Hurwitz determinant is

$$H_n = \begin{vmatrix} a_{n-1} & a_{n-3} & a_{n-5} & \cdot \\ a_n & a_{n-2} & a_{n-4} & \cdot \\ 0 & a_{n-1} & a_{n-3} & \cdot \\ 0 & a_n & a_{n-2} & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{vmatrix}$$

H_k^*, finally, signifies the kth minor determinant of the Hurwitz determinant H_n^* after H_n has been reflected about one of the main diagonals, so that H_n^* reads as follows

$$H_n^* = \begin{vmatrix} a_0 & 0 & 0 & \cdot \\ a_2 & a_1 & a_0 & \cdot \\ a_4 & a_3 & a_2 & \cdot \\ a_6 & a_5 & a_4 & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{vmatrix}$$

Discussion

A. Leonhard.[5] In many cases the linear optimum criterion has shown itself to be very useful if one specifies the additional condition that as many roots of the characteristic equation as possible shall have equal negative real parts. This means, in other words, that the various modes die out with equal time constants. Applying this to the example worked out in the paper, we specify that the linear measure of error shall be a minimum, and that, in addition to the two complex conjugate roots, there shall be a real root equal to the real part of the complex roots. Under those assumptions, we get very practical values for the parameters to be chosen.

The two figures (Fig. 8 of this discussion) which correspond to Fig. 4 of the paper, show the curves designated (l^*), which have been calculated under the conditions mentioned. One can see that the new curves (l^*) lie between those calculated for the practical and the quadratic optimum. The transient response of the output

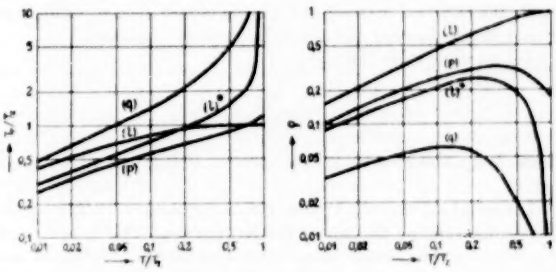


FIG. 8

for $T/T_s = 0.2$ is similar to the one shown for the practical optimum (p) in Fig. 6 of the paper.

As shown elsewhere,[6] the calculation of the parameters according to this method presents no mathematical difficulty, and the amount of calculation that is required is very moderate.

[5] Visiting Professor, Istanbul Teknik Universitè, Istanbul, Turkey. Translated by D. W. Pessen, Minneapolis-Honeywell Regulator Company, Philadelphia, Pa.

[6] "Die selbsttätige Regelung," by A. Leonhard, Springer-Verlag, Berlin, Germany, 1949.

J. M. L. Janssen.[7] In their Statement of the Problem the authors say that the choice of the criterion used to define optimum control action is rather subjective. From a mathematical point of view this statement may be correct. Usually, however, automatic control serves other than mathematical purposes and in industrial applications the criterion for optimum control action is basically an economic one. Thus the criterion is definitely not subjective, although, generally speaking, its formulation in mathematical terms is impracticable. The subjective element comes in because current mathematical formulations aim at mathematical convenience. From this aspect a criterion is practical if it gives rise to simple mathematical operations. This does not necessarily mean that it is also practical from the point of view of the man who must tune a controller.

Now turning to the mathematical side of the paper, some confusion arises from the fact that the transfer functions G_{CL} given in Equations [1*a*] and [1*b*] are not equivalent. Equation [1*a*] relates error to load disturbance. Equation [1*b*], however, does not give an error transfer function. Instead of error it relates system output to set-point disturbance. At first sight, using this transfer function in the quadratic measure of error Q seems incorrect. Fortunately, however, useful results obtain. In order to demonstrate this the discussion in connection with Fig. 3 will be followed.

The curve in this figure certainly applies to set-point disturbances, since the system has a finite steady-state output, and, in accordance with the authors' statement, the disturbance shall be a unit impulse in the time domain. Now for set-point disturbances in a proper control system we have $G_{CL}(0) = 1$. Consequently, there is no freedom to make $G_{CL}(0)$ as small as possible and $G_{CL}(0)$ cannot be a measure of the area enclosed by $G_{CL}(j\omega)$ and the axes. If, on the other hand, the length of the rectangle in Fig. 3 is made as great as possible, then the area, and thus the measure of error Q, is maximized. So in this case the practical optimum is a maximum instead of a minimum. However, Q is not a measure of error in this case. It is rather a measure for the quality of follow-up action. So the maximum found has a real significance.

P. Naslin.[8] The practical example given in the last section confirms the general opinion that the quadratic optimal system is underdamped and the linear optimal system overdamped. Besides, the practical optimum is to all intents and purposes identical with the solution arrived at using standard design techniques based on attenuation and attenuation-phase diagrams.

Let us write G_1 in the form

$$G_1 = \frac{\tau_1 q + 1}{Kq(\tau_2 q + 1)}$$

with $T/T_s = 0.2$, $\tau_1 = T_r/T_s$, $\tau_1/\tau_2 = K/0.2$, and $K = 0.2 + \rho T_r/T_s$.

Let us take K and T_r/T_s as new free parameters; $1/K$ is a gain factor determining the gain at low frequencies. The ratio τ_1/τ_2 of the two time constants being always larger than unity, G_1 appears as the transfer function of an integral controller corrected by a phase-advance system. Then, as K is increased from the value 0.2 corresponding to pure integral control, the gain at low frequencies decreases and the ratio τ_1/τ_2 increases, thus increasing the phase advance. Both of these effects have a stabilizing action. Since, as is apparent from the open-loop gain curve, the pure integral control system is not sufficiently damped, it is clear there must be an optimum combining a good damping and a fast response.

[7] Research Engineer, Royal-Dutch/Shell-Laboratory, Delft, Holland.

[8] Professor, Laboratoire Central de l'Armement, Paris, France.

TABLE 1 VALUES OF PARAMETERS

Setting	K	T_r/T_s	τ_1/τ_2	Notes
q'.....	0.2	0	1	Floating control; underdamped
q.....	0.32	2	3/2	Ineffective; still underdamped
l.....	0.8	1	4	Overdamped
p.....	0.4	0.7	2	OK; Q = 2.5 db
s.....	0.4	1	2	OK; Q = 2 db

Table 1, herewith, gives the value of the parameters corresponding to pure integral control q', quadratic optimum q, linear optimum l, practical optimum p, and standard design techniques s; the corresponding open-loop gain curves are shown in Fig. 9 of this

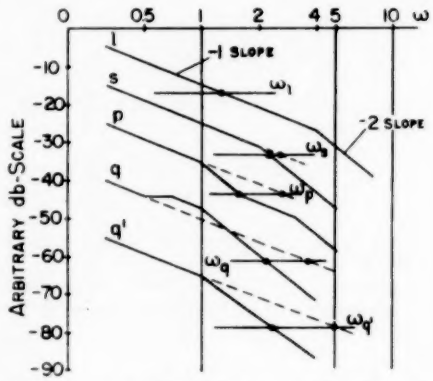


Fig. 9 Open-Loop Gain Curves for Linear Optimum l, Standard Method s, Practical Optimum p, Quadratic Optimum q, and Pure Floating Control q'

(The frequency axis pertaining to each curve is marked ω with the proper index.)

discussion. The q'-setting is underdamped because the gain slope at crossover is —2. The q-setting hardly brings any improvement at all, because the time-constant ratio of the phase-lead system is too small and, besides, the corresponding frequencies are placed in an entirely wrong part of the spectrum. The l-setting is overdamped because the slope at crossover is —1. The p and s-settings are about equivalent, both having a slope of about 3/2 at crossover frequency. Their Black curves, shown in Fig. 10 of this discussion, nearly coincide.

The curves of $|C(p)|$ for the p and s-settings are drawn in Fig. 11. From this viewpoint, the p-setting is slightly better than the s-setting, but the difference is so small as to be irrelevant in practice. Nevertheless, the difference is easily predictable using a remark attributable to Graham.[9]

Within the loop pass-band, we may write approximately

$$|C(q)| \approx 1/|q|\ |G_1(q)|$$

Therefore, if we wish to make this quantity constant within this frequency range, then we must choose G_1 so that it behaves as much as possible like $1/|q|$. It is seen that this condition is slightly better fulfilled for the p-setting than for the s-setting because the first corner frequency of the lead term is slightly higher. But, again, the difference is so small as to be insignificant.

The foregoing gives cause to wonder about the usefulness of optimization criteria. However, one must admit that, if the quadratic optimum, e.g., gives unsatisfactory results in so far as the step response is concerned, it simply means that it is not the correct sort of criterion for the purpose.

On the other hand, in many cases of a continuous random disturbance, the mean-square error does seem quite a fair figure of

[9] "Linear Servo Theory," *Bell System Technical Journal*, October, 1946.

merit. In such a case, do we really care about the shape of the step or impulse response, which will never occur in practice? Nobody seems to know the answer for sure. In any case, the important point is the validity of the criterion. It is not logical to use a criterion and then to grumble about the results it gives. There is absolutely no point, for instance, in using the mean-square criterion and then to decrease the gain somewhat in order to improve the step response.

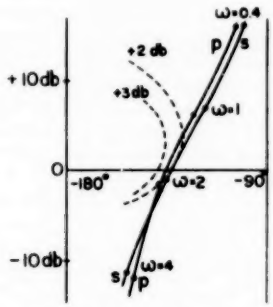


FIG. 10 BLACK CURVES OF p AND s-SETTINGS

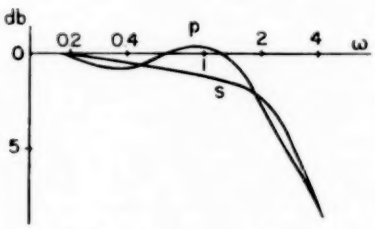


FIG. 11 $|C(q)|$ FOR p AND s- SETTINGS

This does not mean of course that a knowledge of the frequency content of the potential disturbances is not highly desirable. On the contrary, it does seem of advantage, as Graham pointed out, to distribute the error uniformly over the loop pass band and, for that purpose, to make $|G_1|$ as nearly coincident as possible with the disturbance spectrum within the pass band. It should be noticed that this approximation is quite satisfactory if the damping is sufficient ($Q = 2$ to 3 db). The value of the paper under discussion lies in the fact that it seems to give us an easily applicable mathematical method for achieving these results.

The following is only a minor point but on the whole not altogether academic. It does not seem right to say that confinement of the roots to the negative real axis "guarantees an absolutely nonoscillatory . . . . . recovery curve." If G_1 behaves like $1/s$ at low frequencies, then of course $C(0)$ in response to a unit-step disturbance is a constant equal to the algebraic area of the step-response curve, which may in such a case be nonoscillatory, Fig. 12(a), herewith. But, if G_1 behaves like $1/s^2$ at low frequencies, then $C(0) = 0$, which means that the step response, whose net area is now zero, must be oscillatory, Fig. 12(b), even if all the roots are on the negative real axis. Such would be the case, e.g., with

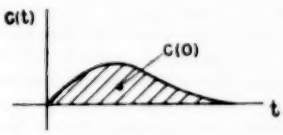
$$C(s) = \frac{2s(5s + 2)}{(s + 1)(2s + 1)(3s + 1)(4s + 1)} = \frac{1}{s + 1} + \frac{2}{2s + 1} + \frac{3}{3s + 1} - \frac{4}{4s + 1}$$

Similarly, in the case of a position-control system, it is not possible to make the responses to a position step nonoscillatory if one wishes the steady-state velocity error to vanish; the velocity error being the integral of the position error, the latter must clearly be oscillatory if the former is to assume a null final value. Such simple relationships should be kept in mind when setting design specifications for a servosystem.

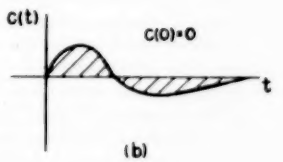
E. F. TROMBLEY.[10] The quadratic measure of error should prove to be the most natural as has been the case in curve fitting by least squares. If we consider all errors $C(t)$ satisfying

$$\int_0^\infty C^2(t)dt < \infty$$

(integration in the sense of Lebesgue) and identify functions equal almost everywhere, then we have a real Hilbert space. Since



(a)



(b)

(a) $G_1(0) \sim \frac{1}{s}$ (b) $G_1(0) \sim \frac{1}{s^2}$

FIG. 12 STEP RESPONSES

Hilbert space theory has been extensively developed, some of this theory should find useful application here and perhaps even suggest further developments.

If we call

$$\frac{1}{\pi}\int_0^\infty |C(j\omega)|^2 d\omega$$

the "quadratic measure of error in the frequency domain," Parseval's formula (Equation [8] of the paper) says that minimizing the quadratic measure of error in the time domain is the same as minimizing the quadratic measure of error in the frequency domain. Hence in so far as the quadratic measure of error is concerned, the time domain and the frequency domain are equivalent. Other measures of error may be considered, such as

$$Q_p = \int_0^\infty |C(t)|^p dt \quad (p > 0).$$

Parseval's formula fails except for $p = 2$. For $p \neq 2$, there seems to be no reasonable correspondence between the time domain and the frequency domain. This indicates that the quadratic measure of error is the only one worth selecting from these possibilities.

If the error also depends on a parameter α and it is assumed that

$$Q(\alpha) = \int_0^\infty \int_0^\infty C^2(\alpha,t)d\alpha\, dt < \infty$$

it should be possible to extend some of the results of this paper to

[10] Dynamicist, Research and Development Department, Pullman-Standard Car Manufacturing Company, Hammond, Ind.

$Q(\alpha)$. Much of the theory of Fourier integrals extends to this case.

The periodic disturbing function is a special case of an almost periodic function and the mean value defined in Equation [17] of the paper is precisely the one used in the study of almost periodic functions. For arbitrary disturbing functions it should be of interest to use summability methods applied to the Fourier integral to pass from the time domain to the frequency domain. The integral mean used in this paper is an example of such a method.

W. W. Seifert.[11] Determination of optimum adjustment is an important aspect of the design of automatic control systems. As the authors indicate, several different optimization criteria are possible. However, they have not given a convincing discussion of their "practical optimum criterion." They seem to make an undue distinction between process-control systems and servomechanisms. This distinction serves merely as a framework to describe the differences in the response of a closed-loop system to input and load disturbances. They discuss at length linear and quadratic optimization criteria and at the end of the paper give a brief description of their practical optimum criterion. The result is that if this criterion has any merit, they have not presented either a convincing mathematical or physical justification for it. At first glance, the results obtained in their practical example appear convincing, but second thought raises doubts whether a fortuitous example was chosen and whether, in general, good results would be obtained.

For the example chosen, the complexity is such that numerical evaluation of the quadratic error can be made easily by substituting typical values for the parameters and by calculation of the error. Examination of the error obtained for several parameter combinations leads quickly to results which are nearly optimum. The authors overestimate the difficulties in applying the quadratic approach.

R. Oldenburger.[12] This paper is particularly timely because of the current interest in the subject of defining and determining optimum controls for physical systems. The particular optimum criteria employed in a given problem will depend to a large extent on the ease with which one can carry through the mathematical theory necessary for the solution of the problem. It is significant that most definitions of optimum controls generally give quite satisfactory results and place one in the desired area of performance.

The Parseval formula (Equation [8] of the paper) can be obtained by fairly simple mathematical manipulations from the theory given by Titchmarsh in his book on the subject.[13] The authors' treatment of the problem of minimizing the mean-square error for the isolated disturbance as well as statistically described disturbances is particularly illuminating, especially in that the results are shown to be completely analogous. Considering the enormous amount of work that must often be carried out to minimize the mean-square error, consideration of simpler and more practical optimum criteria by the authors is a step in the right direction.

If the Conditions [27], [28], and [29] to $\mu = m$ are satisfied it means that the authors are assuming that the function $J(x)$ is of the form

$$J(x) = \frac{N(x)}{G(x) + N(x)} \qquad [36]$$

where the function $G(x)$ is a polynomial whose terms are of higher degree than m. This is a consequence of the linear dependence Relations [30]. If there are not enough parameters to satisfy the conditions mentioned, as is normally the case, the inequalities of Formula [24] are not all fulfilled, and the curve for $J(x)$ may not be monotonically decreasing. If, for example, Equation [27] is satisfied it implies that the curve for $J(x)$ is horizontal at $x = 0$, but it does not imply that the curve will fall thereafter. Clearly, if the curve for $J(x)$ falls too fast the system will be too sluggish, whereas if the curve has too much of a resonance peak the system will very likely be too oscillatory.

There remains the serious problem of proving, if it is so, that the authors are justified in supposing that optimum or nearly optimum control is attained by making as many derivatives of $J(x)$ zero at $x = 0$ as possible. In particular, the following question arises. If $J(x)$ is monotonically nonincreasing, is optimum control achieved by setting as many of the derivatives $J(x)$, $J'(x)$, and so on, equal to zero as possible?

Men will argue forever as to what optimum control performance is. There are very few places, relatively speaking, where everything one wants for an optimum performance can be simultaneously satisfied. The writer has found this possible only in certain nonlinear problems, where minimum overswing, minimum area between the transient curve and the time axis, minimum rms error, minimum return time to equilibrium, and other such desired criteria are all satisfied at the same time.

Harold Chestnut.[14] This is an interesting paper on a much considered problem. The results of the practical approach that the authors suggest, as shown in Figs. 6 and 7, seem to be more satisfactory than the linear or quadratic criteria that have been suggested previously. Although the practical approach produces a desirable transient response, the fact that its spectral density does not have a maximum value at other than zero frequency causes it to depart from the "optimum" servomechanism as actually encountered experimentally. However, the general suitability of its results and its ease of application make it especially attractive to control-system designers in analytical work.

The authors at the outset point out some of the similarities and the differences between process-control systems and servomechanisms. It is worth while to discuss a few of the assumptions that were cited initially to point out that servomechanism requirements may cause criteria, other than the three described by the authors, to be considered for most satisfactory servomechanism settings.

The input to a servomechanism is seldom one which changes slowly compared to the speed of response of the control loop In contrast with process controls where the changes in set point are rare, the servomechanism may be called upon continually to follow a rapidly changing input. One of the principal causes of servo error is caused by its reponse to input motion.

That the process or power-element parameters are predetermined and fixed is generally similar for process control and servomechanisms. However, the extensive use of feedback with a number of supplemental loops in servo elements tends to permit fairly extensive modifications of the inherent power element. As such, its performance, as it finally appears in the control, is appreciably different from its inherent characteristics. Control settings such as these also should be included in establishing an optimum.

In servomechanisms and control systems encountered in practice the number of degrees of freedom and the order of the differential equations for these controls is generally high. The practical method outlined by the authors should simplify greatly the

[11] Assistant Director, Dynamic Analysis and Control Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

[12] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

[13] "Introduction to the Theory of Fourier Integrals," by E. C. Titchmarsh, Clarendon Press, Oxford, England, 1948.

[14] Engineer, General Electric Company, Schenectady, N. Y.

process of obtaining an optimum solution by analytical means, as compared to previous methods.

DUNSTAN GRAHAM[15] and R. C. LATHROP.[16] The authors are to be congratulated on their highly original treatment of an important phase of automatic-control design. Although the synthesis of dynamic systems has received general treatment in the literature, the detailed and systematic investigation of specific criteria for optimum response has been generally overlooked. This is due in large measure to the long-standing lack of a suitable definition for "optimum" response. The "practical optimum" developed in terms of the coefficients of the closed-loop frequency-response function is a significant contribution to the synthesis art.

The authors have convincingly stated the case in favor of the analytic adjustment for optimum results. They quite correctly observed that such a process begins with the definition of a suitable quality criterion and the calculation of the corresponding measure, followed by the determination of those values of the parameters which produce the extreme value of the quality criterion. These two points would seem to emphasize the fact that there are criteria for criteria. A criterion for the response of servomechanisms should have three attributes—reliability, ready applicability, and selectivity.[17, 18] It should select reliably the best response for a variety of systems, should not be difficult to apply in calculations, and should discriminate against systems which are not quite so good as the optimum. The authors of the paper consider and reject two widely known criteria because they fail to meet one or more of these standards.

The linear criterion is unreliable even if the roots of the characteristic equation are restricted to real negative values, since under these conditions, it selects a response which is slower than it need be for most applications. This criterion, however, has a powerful advantage in ease of application. Not only can the criterion value be calculated directly from the Laplace transform of the error response, but also the mathematical conditions for a minimum criterion value (real equal negative roots) are even simpler than the authors have indicated in the paper. If, by a change of time scale, the system characteristic equation is cast into a normalized form so that the coefficients of the highest and lowest powers of the complex variable are both unity, then the remaining parameters (one less than the order of the system) must take the values of the binomial coefficients.[17]

In addition, for the case where the linear control-area criterion is evaluated in terms of the servomechanism error for a unit step function of input, it does not lose all its meaning even if restrictions on the roots are removed, since it continues to signify the velocity coefficient of the servomechanism.[18]

The quadratic criterion is more widely held in high repute. The authors reject it on the basis that the mathematical difficulties involved in the evaluation of the integral vitiate its applicability. To our minds this is not its greatest failing. For simple inputs, the criterion integral may be evaluated easily on an analog computer. Furthermore, tables for the integral have been prepared.[18] The fact that the quadratic criterion may select responses which are unacceptably oscillatory, and that the criterion value is insensitive to fairly large changes in system parameters, leads us to conclude that this measure by itself has relatively little merit.[17]

[15] Chief, Central Synthesis Branch, Wright Air Development Center, Dayton, Ohio.

[16] Chief, Measurements and Analysis Section, Wright Air Development Center.

[17] "The Synthesis of 'Optimum' Transient Response: Criteria and Standard Forms," by D. Graham and R. C. Lathrop, Trans. AIEE, vol. 72, 1953 (technical paper 53-249).

[18] "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., New York, N. Y., 1947.

The writers[17] have reported an investigation of eight different integral criteria which may be applied to servomechanism responses. The criterion which best meets all requirement is the integral of time-multiplied absolute value of error (ITAE), $\int_0^\infty t\,|e|\,dt$. While this criterion is nonanalytic, it is relatively easy to mechanize on an analog computer. It consistently selects a response in accord with intuitive concepts or a priori knowledge of a desirable optimum, and it possesses adequate selectivity.

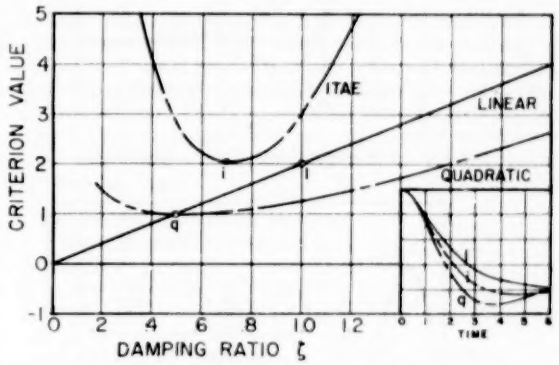


FIG. 13 COMPARISON OF OPTIMIZING CRITERIA FOR A SECOND-ORDER SYSTEM

Fig. 13 of this discussion shows the values of the linear, quadratic, and ITAE criteria as functions of the single parameter ζ, evaluated for a step function of input to the linear second-order system with the normalized transfer function

$$\frac{C(s)}{R(s)} = \frac{1}{s^2 + 2\zeta s + 1} \qquad [37]$$

The inset illustrates the error responses corresponding to the parameter values selected by the various criteria, and identified by the letters *l* for linear, *q* for quadratic, and *i* for ITAE. The slow response of the linear optimum, and the oscillatory response and poor selectivity of the quadratic optimum may be appreciated with the aid of these diagrams. The superiority of the ITAE criterion is even more evident on progressing to high-ordered systems.

It is pertinent to note that in the case of the second-order unit-numerator system, the practical optimum of the authors selects a parameter identical in value to the one chosen by the ITAE criterion. For the higher-ordered unit-numerator systems, the practical optimum selects parameter values apparently identical to those of the Butterworth filters.[17] This fact is certainly in accord with the concepts from which the practical optimum was developed, for the Butterworth filters are characterized by an attenuation curve flat to the highest possible frequency, and decreasing monotonically. The corresponding phase angles, however, are large in the vicinity of the natural frequency, especially for higher-ordered systems, and the transient responses of the higher-ordered Butterworth filters are poorly damped. The error responses of fourth-order unit-numerator systems, to unit initial conditions or step functions of input, are illustrated in Fig. 14, herewith. The ITAE criterion in this and similar cases of still higher order selects a response which is much faster than the linear optimum and yet is better damped than the practical optimum.

If an analog computer is available, the ITAE criterion is commended for the synthesis of optimum control systems. The simplicity and power of its use may be illustrated by the fact that the example problem given in the paper was set up and solved

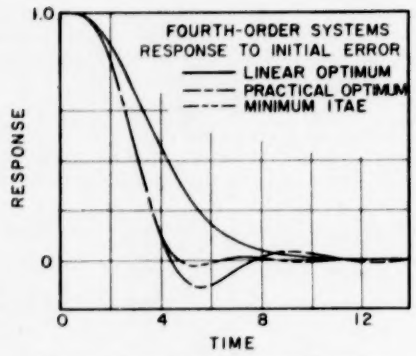


FIG. 14 FOURTH-ORDER SYSTEM RESPONSE

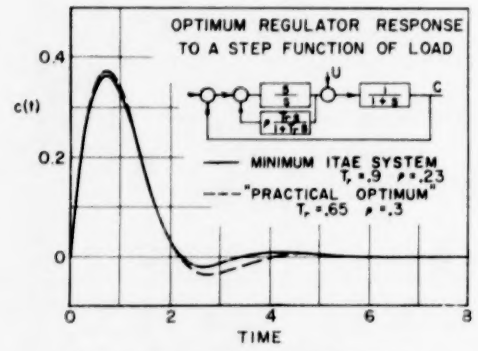


FIG. 15 OPTIMUM REGULATOR RESPONSE

thirteen times to obtain a minimum criterion value in about 1 hr. The result is illustrated in Fig. 15 of this discussion which is directly comparable to Fig. 6 of the paper. The minimum ITAE response may be seen to have a slight advantage in quality of control over the practical optimum. The advantage is so slight, however, that without an analog computer or "standard forms"[17] for the transfer function based on the ITAE criterion, the designer would seem to be well advised to employ the practical optimum as a valid and not too difficult approach to the optimum adjustment of control loops.

AUTHORS' CLOSURE

The method mentioned by Professor Leonhard may yield quite useful results in many cases, as, for example, when applied to the simple numerical example. In the opinion of the authors, however, this method is not an optimizing procedure in a strict sense. The characteristic equation is based only upon the homogeneous part of the differential equation describing the system under investigation. The characteristic equation, therefore, can only provide information concerning the roots of the system, but not concerning the specific initial conditions, and certainly not concerning the disturbance function. On the other hand, among all the coefficients in the numerator of the transient-response Laplace transform, only the zero-order term (b_0 in Equation [11]) shows up in the linear control area, while all the other coefficients are disregarded. These coefficients contain the specific initial conditions of the system, and, in general, are also modified by the disturbance function. A true optimizing procedure, however, must take the initial conditions as well as the disturbance function into account, because the transient recovery curve is affected by both. For example, a control system must basically be adjusted differently, depending on whether its most important function is to counteract load disturbances (as is the case in process control) or to respond accurately to set-point changes (as is the case with servomechanisms). Neither the characteristic equation nor the linear control area can take these important factors properly into account.

Mr. Janssen points out that every optimizing problem is, in the end, based on economic considerations. The authors subscribe to this conclusion as completely as they do to the view that starting with the foregoing standpoint in mind, there can be no subjective differences of opinion concerning control criteria. However, the question as to what type of quality deviations are least harmful economically can hardly be formulated mathematically in a general manner. This is especially so because the fixed costs of the plant and of the control installation play a deciding role in the whole question. If we limit ourselves (as is done in this paper) to the considerably narrower problem of adjusting a given controller to a given process, than the criteria which apply can be quite different from case to case due to economic considerations. The authors' statement that the choice of criterion is rather subjective is to be understood in this sense, although, undoubtedly, it would have been better to say that the choice of criterion is "different from case to case."

Mr. Janssen's comment that the two transfer functions given by Equations [1*a*] and [1*b*] are not equivalent is not valid, in the opinion of the authors. The transfer function [1*a*] relates system output or controlled variable to a load disturbance at the input of the process, while the transfer function [1*b*] relates system output to a set-point disturbance at the input of the controller. Mr. Janssen is correct, however, in that the transfer function [1*b*] is not directly suitable for determining the adjustment of servomechanisms by means of the method described in the paper.

In process control, the deviations of the controlled variable which are due to a load disturbance must be made to disappear as quickly and as completely as possible. With servomechanisms, however, the problem consists of having the controlled variable follow changes of the set point as quickly and completely as possible. These two problems, which are quite different, become equivalent if, for the servomechanism, we use the transfer function relating the error or deviation to the set-point disturbance, given by the equation

$$G_{CL} = 1 - \frac{G_1 G_2}{1 + G_1 G_2} = \frac{1}{1 + G_1 G_2} ,$$

This error, similarly to the deviation caused by a load disturbance in process control, must be made to disappear as quickly and as completely as possible. In this way the described method to find proper controller settings is applicable for both types of disturbances. The authors thank Mr. Janssen for calling attention to this important point.

The authors thank Professor Naslin for his valuable comments and agree with his point that confinement of the roots to the negative real axis does not necessarily guarantee a nonoscillatory response.

The authors cannot agree with the view expressed by Mr. Seifert. On the contrary, they believe that they have pointed out expressly that there are no basic differences in the treatment of process-control systems and of servomechanisms. It must be admitted that the numerical example chosen might have been made too simple for the sake of clarity and a compact presentation. This was done in order to make the advantages and the capabilities of the suggested practical optimum really clear. The experience of the authors has shown, however, that the convenience and practicability of the procedure is even more pronounced for more complicated problems It cannot be disputed that the

numerical evaluation of any given case can always be carried out using the quadratic optimum, sometimes even with a moderate amount of effort. However, a general solution and an investigation of the relations controlling the free parameters is certainly simpler and less time-consuming when the practical optimum is used.

The doubts expressed by Dr. Oldenburger for the case that a sufficient number of parameters are not available for the optimizing procedure are completely justified from a mathematical standpoint. Especially in the field of process control, however, the controlled systems occurring in practice usually consist of pure time lags. Assuming common disturbing functions, these time lags contribute only a falling characteristic to the frequency spectrum of the closed loop. If, therefore, the frequency spectrum contains resonance peaks, then these must somehow be produced by the controller. However, controllers whose properties produce resonance peaks should also provide the possibility for adjusting these properties to eliminate resonance effects. If this is not possible (for example, because of an insufficient number or an insufficient range of controller adjustments), then the controllers are either deficient in their design or are not being used in a correct manner. The suggested procedure has the special advantage, that, in cases such as the foregoing, it will clearly point out deficient assumptions of this sort and, at the same time, will point toward the way of possible improvements. This should also be the case with less desirable controlled systems containing mass effects, as they often appear in servomechanisms.

In summary, the authors are thankful for the interest shown their work, and for all discussions, including those which were not specifically answered in this closure. The methods suggested in the paper grew out of the desire to supply the planning engineer with a clear and simple method for the investigation and evaluation of control loops. The authors realize that their proposals are still in need of supplementation and completion in various respects. They hope, however, that their work has provided useful and stimulating ideas

Recent Advances in Nonlinear Servo Theory

By J. M. LOEB,[1] PARIS, FRANCE

This paper concerns the so-called "filtered nonlinear servos," i.e., servos where the signal undergoes a filtering effect from a linear part of the device. "Hereditary phenomena," such as Coulomb friction, gear backlash, and so on, are dealt with. A general stability criterion is derived, and it is shown how the Nyquist criterion appears as a particular case. The criterion indicates whether or not a servo is liable to hunt, but it is still necessary to state whether or not the possible hunting will have a limited amplitude. An approach to this problem is sketched and it is shown how the first results may be considered as quantitatively valid for quasi-linear systems. Some conclusions previously derived by assuming the transfer functions to be analytic appear as a limiting case of the present results.

Introduction

RECENT works have shown that, once the linear servo theories have been developed, it becomes necessary to deal with the case of real, i.e., nonlinear, servomechanisms. In France, J. R. Dutilh (1, 2)[2] and in the United States, Kochenburger (3) have studied contactor servos, apparently independently. According to their conceptions, analysis is made possible by a filtering effect, coming from inert elements such as the rotor of the electric motor. This allows one to overlook all Fourier harmonics.

This conception proved very useful, for at the same time, J. R. Dutilh indicated a stabilizing device that permits the construction of contactor servos with the same practical performance as so-called linear servos. The devices studied by Dutilh and Kochenburger correspond to the diagram, Fig. 1, where the

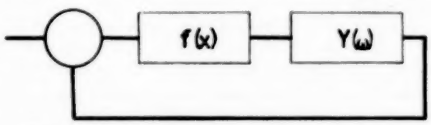


FIG. 1 BLOCK DIAGRAM OF AN ON–OFF SERVO

first element has a transfer function $f(x)$ depending on amplitude, but independent of frequency, and the second element is a conventional linear one characterized by a transfer function $y(\omega)$.

It is the aim of the present paper to extend the known results in the following directions: (*a*) Hereditary phenomena; (*b*) more general nonseparable mechanisms; (*c*) connection with Nyquist's criterion; (*d*) stability of possible hunting.

Hereditary Phenomena

The expression "hereditary phenomena" was coined by Vito Volterra, and applies to phenomena in which the output is not a function of the input but a functional. Hysteresis phenomena (e.g., in amplidynes), gear backlash, Coulomb friction, and so on, are hereditary phenomena.

Instead of trying to express the output y as a function of an infinite number of variables, i.e., all the values of the input x, it is at the same time more convenient and more general to investigate the properties of the functional by examining the time variation of y when x is a function of time. Stability is one of the concerns of the servo engineer, who will try to analyze hunting in which $x(t)$ and therefore $y(t)$ are periodic time functions. The usual functionals are sufficiently and easily described by assuming that x undergoes a cyclic variation.

The well-known magnetic hysteresis pattern is found in the following phenomena:

Coulomb Friction. Authors who deal with "real" servos are prone to put into their calculations friction forces of the viscous type (proportional to velocity) because these yield linear equations.

Resistance resulting from dry friction, or "Coulomb friction" is often encountered in practice. Let us consider, for instance, a body of weight P lying on a plane with a friction coefficient f. The friction is given by $-f\ P\ \mathrm{sgn}\ (v)$ where v is the velocity. Here "sgn (v)" means the sign of v.

Coulomb friction is generally associated with forces of inertia as in the following example: A rotor of moment of inertia J, with a friction torque C_0, is acted upon by torque Γ which is a periodic function of time. Let us calculate the torque Γ that gives the rotor an angular velocity which is a sinusoidal function of time (curve I, Fig. 2*a*). To the torque $J\dot{\omega}$ (curve II) which cancels the inertia forces is added a constant torque C_0 which has the sign of ω, the sum of them being shown on curve III. The curve in Fig. 2(*b*), akin to the hysteresis curve, gives the torque Γ plotted against $J\dot{\omega}$. Points $A\ B\ C_1\ C_2\ D\ E\ F_1\ F_2$ of both curves correspond.

If, instead of assuming $J\dot{\omega}$ to be known, we want to calculate it from Γ, this latter would be represented by a continuous curve and the angular acceleration by a discontinuous one.

Gear Backlash. The curves are obtained in a similar way. If x is the angular position of the driving gear, and y the angular position of the driven gear, the values of y plotted against x are given by the curve *A-B-C-D* in Fig. 3 where 2σ is the backlash.

This result is valid only when the cycle is going on. The beginning of the motion, may in effect take place when the gears are already in contact and the curve begins at the segment Oa. It is in this way that we may introduce Vito Volterra's assumption of independence from old states.

Analysis of Nonseparable Systems

This analysis, the main features of which have been published by the author of this paper (4, 5), is applicable, as in the case of the two preceding cases, to filtered systems but this is the only assumption that limits its field of application. It includes indeed all the causes of nonlinearity (thresholds, saturation, hereditary phenomena, and so on).

Criterion of Set of Curves. The open-loop system may be represented by a transfer function $A(x, \omega)$ defined by its gain G and its phase Φ, both of which depend on the frequency ω and amplitude x of a sinusoidal input. This function is plotted in polar co-ordinates G, Φ, by a double family of curves x = const and

---

[1] Associate Director, Centre National d'Etudes des Télécommunications.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 18, 1953. Paper No. 53—A-19.

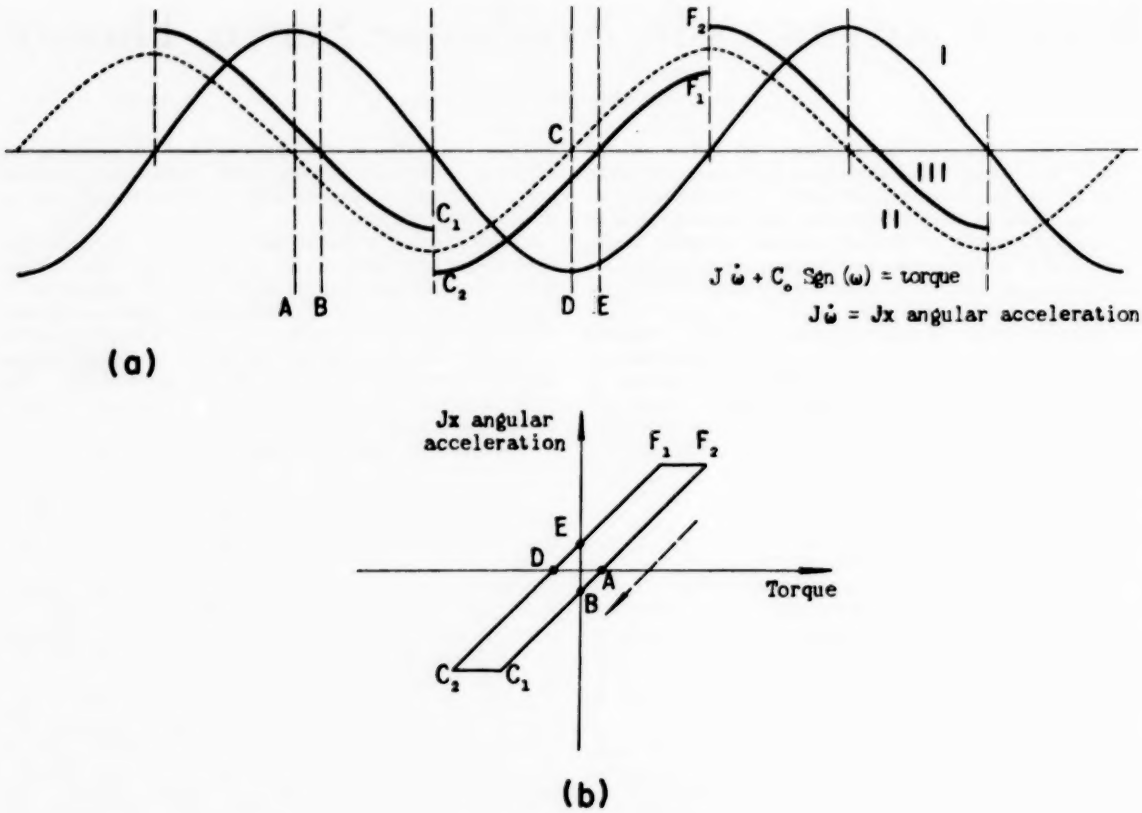


FIG. 2 COULOMB FRICTION
(*a*, Time functions. *b*, Functional curve.)

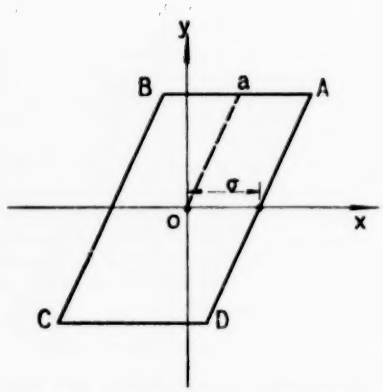


FIG. 3 FUNCTIONAL CURVE OF BACKLASH

ω = const, Figs. 4(*a* and *b*). This set generally covers a region R of the plane. If this region includes the point N where $G = 1$, $\Phi = \pi$, there exists a pair of values x_0 and ω_0 for x and ω such that

$$1 + A(x_0, \omega_0) = 0$$

An oscillation of constant amplitude x_0 and angular frequency ω_0 is then in agreement with the laws of the motion. This does not yet mean that the oscillation actually will take place. We will still have to formulate the stability criterion of this oscillation, and it is then that the difficulties begin.

The first criterion is stated as follows: Plot the family of curves representing the point

$$G(x, \omega), \Phi(x, \omega)$$

in polar co-ordinates. If the region occupied by this family does not include the point N ($G = 1$, $\Phi = \pi$) no oscillation can take place, and the servomechanism is stable. If the region occupied by the set includes the point N, two conclusions are possible; i.e., the operation x_0, ω_0 is, or is not stable.

One must point out that the set of curves may cover several times a part of the plane with different values of x and ω. Riemann surfaces must then be used.

Back to Nyquist's Criterion. All physical systems are amplitude-limited and it is well known that the exponential increase of oscillations, as foreseen by the linear theory, is actually limited by saturation. Saturation causes the modulus G of the transfer function to be a decreasing function of x. As a matter of fact, when saturation takes place the output is limited, no matter how large the input. So the ratio G of their amplitudes is decreasing, and tends toward zero when x increases.

For instance, if complete saturation is going on, i.e., if y is a constant, $f(x)$ is directly proportional to $1/x$. Figs. 4(*a* and *b*)

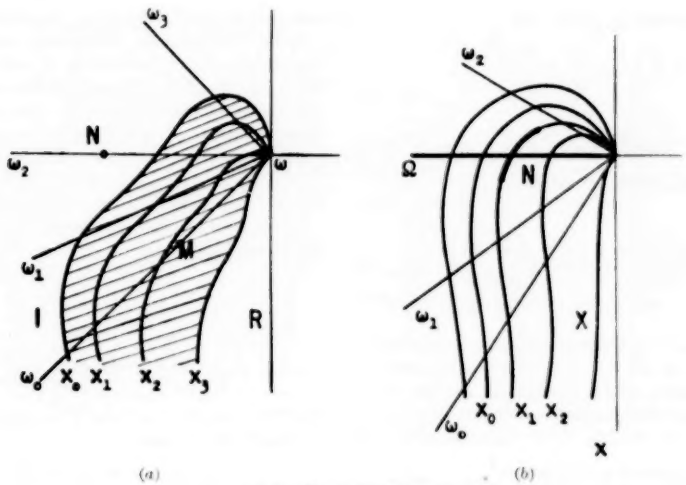


FIG. 4 TYPES OF SERVOS

(*a*, Stable servo. *b*, Servo liable to hunt.)

indicate the set of curves drawn following our criterion in the case of a servomechanism that comprises the following:

A nonlinear amplifier (with saturation, but without any threshold effect).

A motor, the inductance of which cannot be overlooked. In Fig. 4(*a*), where the servo is assumed to be stabilized with an antihunt network, the family of curves is limited by the ordinary Nyquist curve I plotted x_0, because, in the absence of a threshold effect, the modulus G is generally a decreasing function of x, and is maximum in the linear case, i.e., when x is small. Then all the curves of the family are located between I and the negative imaginary axis. The region occupied by the family of curves is actually limited by the "negative" imaginary axis, because the curve I (linear case) passes through the point O for $\omega = \infty$, and because, for large values of x, the curves, such as x_3, come nearer and nearer to the negative imaginary axis.

If the antihunt device is well adjusted, the point N lies outside the region R and the servo is stable.

In Fig. 4(*b*) no antihunt device has been provided, the family of curves contains one pair of curves for $x = x_0$, $\omega = \omega_0$ which pass through N and the servo can go on hunting provided this oscillation be stable. As we will see later, it is indeed possible that a pair of values x_0 and ω_0 will give $G = 1$ and $\Phi = \pi$ but that, however, this hunting operation cannot be sustained. This can happen when a slight disturbance causes the system to go back toward rest or toward another hunting operation. If, after such a disturbance, the system comes back toward the same hunting operation, this latter will be called "stable." In the opposite case, it will be called "unstable." In this case the application of our stability criterion gives the same results as Nyquist's with a further result that we are told what will really happen when nonlinear effects take place and when self-sustained oscillations are going on.

Conditional Stability. Following our theory, this is a case of metastability. There exists an amplitude X_0 of oscillation such that when $x > X_0$ the gain is reduced in such a way that hunting can begin. As long as the amplitude x of oscillation is smaller than X_0, no stable hunting is possible.

STABILITY OF OSCILLATIONS—"TRANSIENT" THEORY

We must now examine the problem of stability of possible oscillations, and in order to do so, we have to calculate how the characteristics of the motion (x and ω) are varying spontaneously after the system has been disturbed slightly from its oscillating operation, and the disturbance has been removed. We will follow a method defined by G. Cahen (6) in a study devoted to the "pulling effect," and also dealt with in a paper by G. Cahen and J. M. Loeb (7).

A First Approach. In order to get the approach well understood, we will first quote a very crude theory of the author (5) which later will appear as valid in a limiting case.

We must now determine the behavior of a feedback system in the neighborhood of the operation defined by an amplitude x_0 and an angular frequency ω_0 such that

$$1 + A(x_0, \omega_0) = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [1]$$

The analysis must involve the time derivative $\dot{x}$ of the amplitude x. (Throughout the paper the dot means derivative with respect to time.) It will be convenient to introduce this unknown through the relative increment, $\tau = \dot{x}/x$. This quantity will be considered as the imaginary part of a complex angular frequency $-j\tau + \omega$. The evolving system is characterized by an angular frequency $\omega_0 + \epsilon$, a relative increment τ (τ is negative when the system evolves toward the x_0, ω_0 operation), and an amplitude $x_0 + \xi$.

These quantities are subject to the general equation of feedback systems

$$1 + A(X_0 + \xi, -\tau j + \omega_0 + \epsilon) = 0 \ldots\ldots\ldots [2]$$

This complex equation gives two relationships between ξ, τ, and ϵ. By eliminating ϵ one gets the desired relationship between τ and ξ.

As we limit our investigation to the operations that are very near to the operation defined by x_0 and ω_0, we may consider ξ, τ, and ϵ as very small quantities.

If we expand Equation [2] in powers of ξ, τ, and ϵ, if we limit the expansion to the first terms, and if we take Equation [1] into consideration, we write

$$\xi \frac{\partial A}{\partial x} + (-j\tau + \epsilon) \frac{\partial A}{\partial \omega} = 0 \ldots\ldots\ldots\ldots [3]$$

When $A = U + jV$ Equation [3] gives

$$\left.\begin{aligned} \xi\frac{\partial U}{\partial x} + \tau\frac{\partial V}{\partial \omega} + \epsilon\frac{\partial U}{\partial \omega} &= 0 \\ \xi\frac{\partial V}{\partial x} - \tau\frac{\partial U}{\partial \omega} + \epsilon\frac{\partial V}{\partial \omega} &= 0 \end{aligned}\right\} \quad \ldots\ldots\ldots [4]$$

By eliminating ϵ one gets the desired relationship

$$\left[\left(\frac{\partial U}{\partial \omega}\right)^2 + \left(\frac{\partial V}{\partial \omega}\right)^2\right]\tau = -\xi\begin{vmatrix} \frac{\partial U}{\partial x} & \frac{\partial V}{\partial x} \\ \frac{\partial U}{\partial \omega} & \frac{\partial V}{\partial \omega} \end{vmatrix} \quad \ldots\ldots [5]$$

If the Jacobian determinant in the second member is positive, τ presents an opposite sign to the one of ξ, which means that, if the system undergoes a slight disturbance, it comes back to the oscillating operation defined by x_0 and ω_0. According to our definition, this operation is stable. It is unstable when the Jacobian is negative.

Objections. This approach is liable to three major objections:

1 When writing the partial derivative $\partial A/\partial\omega$, one assumes that A is an analytic function of the complex variable $\omega - j\tau$.

This is not true, in general, even when the system consists of a nonlinear element whose transfer function does not depend on frequency, and a conventional linear element.

2 One neglects the fact that A may depend not only on τ but equally on the derivative $\dot{\epsilon}$.

Indeed, we will see later that the notion of "instantaneous frequency" is rather misleading and that the response of a four-terminal network to a frequency-modulated input cannot be derived by means of the instantaneous frequency without any correction. We must point out that the angular frequency does not remain constant when the system is evolving between the disturbed operation and the steady operation defined by x_0 and ω_0, and that it is necessary to apply the correct formulas relative to frequency-modulated inputs (see Carson and Fry, reference 8).

3 It is not proved that the approach including first derivatives ξ and $\dot{\epsilon}$ is sufficient.

Rigorous Approach. Such an approach would consist in:

1 Defining the system by the modulus G (gain) and the phase angle Φ of the transfer function, without assuming any analyticity conditions.

2 Writing that G and Φ depend not only on ξ, ϵ, and τ, but on higher derivatives of these latter quantities with respect to time, and not on time explicitly.

A system operating in the neighborhood of the operation defined by x_0 and ω_0 would then be subject to the following equations

$$\left.\begin{aligned} G\left(x_0 + \xi,\ \omega_0 + \epsilon,\ \frac{\dot{\xi}}{x_0},\ \dot{\epsilon}\,\ddot{\xi}\ldots\right) &= 1 \\ \Phi\left(x_0 + \xi,\ \omega_0 + \epsilon,\ \frac{\dot{\xi}}{x_0},\ \dot{\epsilon},\ddot{\xi}\ldots\right) &= \pi \end{aligned}\right\} \quad \ldots\ldots [6]$$

By expanding and eliminating ϵ, one would arrive at a differential equation of a high degree (this degree would be limited only by the approximation desired) such as

$$\xi + K_1\dot{\xi} + K_2\ddot{\xi} + K_3\dddot{\xi} + \ldots = 0 \quad \ldots\ldots [7]$$

In the general case it would be simply impossible to carry out such calculations. However, we can state the following remark:

If the coefficient K_1 is negative, one may say with certainty that the operation defined by x_0 and ω_0 is unstable, because a slight positive disturbance ξ causes a further increase of ξ.

If this coefficient is positive, one cannot state that the operation is stable. It is not proved, in such a case, that x will not begin to oscillate with increasing amplitudes.

First-Order Approach. The sign of the coefficient K_1 of Equation [7] will be determined by the following approach:

As we limit our investigation to the first derivatives of ξ and ϵ, Equations [6] become

$$\left.\begin{aligned} G\left(x_0 + \xi,\ \omega_0 + \epsilon,\ \frac{\dot{\xi}}{x_0},\ \dot{\epsilon}\right) &= 1 \\ \Phi\left(x_0 + \xi,\ \omega_0 + \epsilon,\ \frac{\dot{\xi}}{x_0},\ \dot{\epsilon}\right) &= \pi \end{aligned}\right\} \quad \ldots\ldots [8]$$

With $G(x_0, \omega_0, 0, 0) = 1$ and $\Phi(x_0, \omega_0, 0, 0) = \pi$, a series expansion of Equations [8] yields

$$\left.\begin{aligned} \xi\frac{\partial G}{\partial x} + \epsilon\frac{\partial G}{\partial \omega} + \frac{\dot{\xi}}{x_0}\frac{\partial G}{\partial \tau} + \dot{\epsilon}\frac{\partial G}{\partial \dot{\omega}} &= 0 \\ \xi\frac{\partial \phi}{\partial x} + \epsilon\frac{\partial \phi}{\partial \omega} + \frac{\dot{\xi}}{x_0}\frac{\partial \phi}{\partial \tau} + \dot{\epsilon}\frac{\partial \phi}{\partial \dot{\omega}} &= 0 \end{aligned}\right\} \quad \ldots\ldots [9]$$

where we have replaced dx, $d\omega$, $d\tau$, and $d\dot{\omega}$ by ξ, ϵ, τ, $\dot{\epsilon}$, respectively. We must now draw from Equation [9] the desired relationship between ξ and $\dot{\xi}$. This can be done by eliminating ϵ and $\dot{\epsilon}$ from Equations [9] and a third equation obtained by differentiation.

Elimination of ϵ gives (by multiplying the first by $\partial\Phi/\partial\omega$, the second by $-\partial G/\partial\omega$ and adding)

$$\xi\left(\frac{\partial G}{\partial x}\frac{\partial \Phi}{\partial \omega} - \frac{\partial \Phi}{\partial x}\frac{\partial G}{\partial \omega}\right) + \frac{\dot{\xi}}{x_0}\left(\frac{\partial G}{\partial \tau}\frac{\partial \Phi}{\partial \omega} - \frac{\partial \Phi}{\partial \tau}\frac{\partial G}{\partial \omega}\right) + \dot{\epsilon}\left(\frac{\partial \Phi}{\partial \omega}\frac{\partial G}{\partial \dot{\omega}} - \frac{\partial G}{\partial \omega}\frac{\partial \Phi}{\partial \dot{\omega}}\right) = 0 \quad [9a]$$

In the same way, multiplying the first equation by $\partial\Phi/\partial\dot{\omega}$, the second by $-\partial G/\partial\dot{\omega}$ and adding, we get

$$-\epsilon\left(\frac{\partial G}{\partial \dot{\omega}}\frac{\partial \Phi}{\partial \omega} - \frac{\partial \Phi}{\partial \dot{\omega}}\frac{\partial G}{\partial \omega}\right) + \xi\left(\frac{\partial G}{\partial x}\frac{\partial \Phi}{\partial \omega} - \frac{\partial \Phi}{\partial x}\frac{\partial G}{\partial \dot{\omega}}\right) + \frac{\dot{\xi}}{x_0}\left(\frac{\partial G}{\partial \tau}\frac{\partial \Phi}{\partial \dot{\omega}} - \frac{\partial \Phi}{\partial \tau}\frac{\partial G}{\partial \dot{\omega}}\right) = 0 \quad [9b]$$

Taking the derivative of this last equation, and neglecting $\ddot{\xi}$ and other terms, we get

$$\dot{\epsilon}\left(\frac{\partial G}{\partial \omega}\frac{\partial \Phi}{\partial \omega} - \frac{\partial \Phi}{\partial \dot{\omega}}\frac{\partial G}{\partial \omega}\right) = \dot{\xi}\left(\frac{\partial G}{\partial x}\frac{\partial \Phi}{\partial \dot{\omega}} - \frac{\partial \Phi}{\partial x}\frac{\partial G}{\partial \dot{\omega}}\right) \quad [9c]$$

Whence by replacing, in Equation [9a], $\dot{\epsilon}$ by its value obtained from Equation [9c] we get equation

$$\xi\left(\frac{\partial G}{\partial x}\frac{\partial \phi}{\partial \omega} - \frac{\partial \phi}{\partial x}\frac{\partial G}{\partial \omega}\right) + \frac{\dot{\xi}}{x_0}\left[\left(\frac{\partial G}{\partial \tau}\frac{\partial \phi}{\partial \omega} - \frac{\partial \phi}{\partial \tau}\frac{\partial G}{\partial \omega}\right) + x_0\left(\frac{\partial G}{\partial x}\frac{\partial \phi}{\partial \dot{\omega}} - \frac{\partial \phi}{\partial x}\frac{\partial G}{\partial \dot{\omega}}\right)\right] = 0 \quad [10]$$

The writing of these relations is made simpler by introducing

a, b, c, d, α, β, γ, and δ so that

$$\left.\begin{aligned} \frac{1}{G}\frac{\partial G}{\partial x} &= a \sin \alpha, \quad \frac{1}{G}\frac{\partial G}{\partial \omega} = b \sin \beta, \quad \frac{1}{G}\frac{\partial G}{\partial \tau} = x_0 c \sin \gamma, \quad \frac{1}{G}\frac{\partial G}{\partial \dot{\omega}} = d \sin \delta \\ \frac{\partial \phi}{\partial x} &= a \cos \alpha, \quad \frac{\partial \phi}{\partial \omega} = b \cos \beta, \quad \frac{\partial \phi}{\partial \tau} = x_0 c \cos \gamma, \quad \frac{\partial \phi}{\partial \dot{\omega}} = d \cos \delta \end{aligned}\right\} \quad [11]$$

As an example, if one considers a linear four-terminal network the transfer function g / φ is analytic. So is its logarithm which is $\log g + j\varphi$ and the well-known relationship that characterizes analytic functions is written

$$\frac{\partial}{\partial \tau} \log g = \frac{\partial \varphi}{\partial \omega} \quad \text{and} \quad \frac{\partial}{\partial \omega} \log g = -\frac{\partial \varphi}{\partial \tau}$$

Hence

$$\frac{\partial g}{\partial \tau} = \frac{\partial \varphi}{\partial \omega}, \quad \frac{1}{g}\frac{\partial g}{\partial \omega} = -\frac{\partial \varphi}{\partial \tau} \quad \ldots\ldots\ldots\ldots [12]$$

From Equations [11] we derive

$$\tan \beta = \frac{\frac{1}{g}\frac{\partial g}{\partial \omega}}{\frac{\partial \varphi}{\partial \omega}} = -\frac{\frac{\partial \varphi}{\partial \tau}}{\frac{1}{g}\frac{\partial g}{\partial \tau}} = -\cot \gamma. \quad \ldots\ldots [13]$$

Then

$$\gamma - \beta = \frac{\pi}{2}$$

Returning to the nonlinear case, and limiting the investigation to the first-order equation (i.e., assuming that $\dot{\epsilon}$ and τ are constant) we may write

$$\frac{\dot{\xi}}{\xi} = -\frac{1}{K_1} = -\frac{ab \sin (\alpha - \beta)}{bc \sin (\gamma - \beta) + ad \sin (\alpha - \delta)} \quad [14]$$

When K_1 is negative the operation is unstable, whereas when K_1 is positive the operation is stable.

We may note that, even if $\ddot{\xi}$ cannot be overlooked, the Condition [14] is still valid in the case of unstable operation. In this case the coefficients of ξ and $\dot{\xi}$ are of different signs, and Equation [7] has certainly one or several roots in the right half plane. However, there is a difference in the character of instability owing to the signs of the coefficients K_1, K_2, K_3. If K_1 is negative, it means that as soon as the first order is considered, a slight increase ξ causes the system to depart from the state of operation $\xi = 0$. The system behaves in this case like a servo that is assembled in the wrong way, i.e., with the incorrect feedback sense. If K_1 is positive, that does not actually mean that the state of operation $\xi = 0$ is stable, but the system behaves like a servo having the correct feedback sense, but which might be dynamically unstable. This may be considered as an initial justification for the simplification admitted in this paragraph.

Case of Separable Systems

We will now investigate in more detail an application of the method to more usual forms of nonlinear servos.

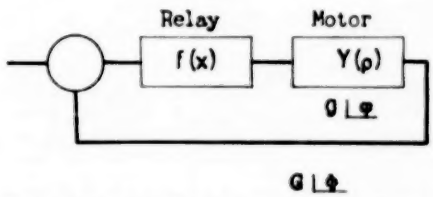
Separable Servos. We will now deal with systems that consist of a nonlinear element (relay, saturated amplifier, and so on) whose transfer function $f(x)$ may introduce phase shift, but is only amplitude-dependent, and which is called the "saturator;" also a filtering linear element, whose transfer function is only frequency-dependent. Its gain g and its phase φ meet the analyticity Conditions [12]. *This will not be the case of the whole transfer function:* $G/\underline{\Phi} = f(x)\ g/\underline{\varphi}$ although $f(x)$ does not depend on ω and τ. This is due to the fact that $f(x)$ modifies the value of the relative increment τ.

Case Where Saturator Precedes Linear Network, Fig. 5

Let y be the output of the saturator corresponding to the input x, where $y = x f(x)$. Then

$$\tau_1 = \frac{\dot{y}}{y} = \frac{\dot{x}}{x}\left(1 + x\frac{f'(x)}{f(x)}\right) = \tau\left(1 + \frac{xf'}{f}\right) \quad \ldots\ldots [15]$$

where f' is the derivative with respect to x.



G/Φ

Fig. 5 Transfer Functions of Separate Parts

Partial derivatives relative to τ (the relative increment of x) are used in Equation [12] and not those relative to τ_1 (the increment of y). For any function F

$$\frac{\partial F}{\partial \tau} = \frac{\partial F}{\partial \tau_1}\left(1 + \frac{xf'}{f}\right) \quad \ldots\ldots [16]$$

Remembering that Equation [12] must be applied to τ_1 and assuming now that $f(x)$ causes no phase shift, we obtain the first six derivatives

$$\left.\begin{aligned} \frac{1}{G}\frac{\partial G}{\partial x} &= \frac{f'}{f}, \quad \frac{1}{G}\frac{\partial G}{\partial \omega} = \frac{1}{g}\frac{\partial g}{\partial \omega}, \quad \frac{1}{G}\frac{\partial G}{\partial \tau} = \frac{\partial \varphi}{\partial \omega}\left(1 + x\frac{f'}{f}\right) \\ \frac{\partial \phi}{\partial x} &= 0, \quad \frac{\partial \phi}{\partial \omega} = \frac{\partial \varphi}{\partial \omega}, \quad \frac{\partial \phi}{\partial \tau} = -\frac{1}{g}\frac{\partial g}{\partial \omega}\left(1 + \frac{xf'}{f}\right) \end{aligned}\right\} \quad [17]$$

All these coefficients can be derived from the knowledge of $f(x)$ and of the transfer function of the linear network.

Calculation of $\partial g/\partial \dot{\omega}$ and $\partial \varphi/\partial \dot{\omega}$. Carson and Fry (8) have shown that if a linear four-terminal network has a transfer function $\gamma(s)$, its response to the input

$$x = x_0 e^{j(\omega t + \frac{1}{2}\dot{\omega}t^2)}$$

is not given by replacing s by $j(\omega_0 + \dot{\omega}t)$.

There is a correcting term ΔY, and we will give only the first term of its expansion

$$\Delta Y = j\frac{\dot{\omega}}{2}\frac{d^2Y}{ds^2} \quad \ldots\ldots [18]$$

Here d^2Y/ds^2 is not an operator; it is the value of the second derivative of $Y(s)$ when $s = j(\omega_0 + \dot{\omega}t)$.

If we assume that $\dot{\omega} \ll \omega_0^2$ and call y the output, the correct answer will be

$$\frac{y}{x} = Y(s) + \Delta Y(s)$$

Here s is the variable $j(\omega_0 + \dot{\omega}t)$. For small values of t, i.e., when the system is operating near the operation (x_0, ω_0), the correcting term $\Delta Y(s)$ may be written as $\Delta Y(\omega_0 j)$, while the principal term Y must still be written

$$Y[j(\omega_0 + \epsilon)]$$

If $d^2Y/ds^2 = Z + Wj$ we get the coefficients to be used in Equations [10] and [11], by the following derivation:

From Equation [18] we have

$$\Delta Y = j\frac{\dot{\omega}}{2}(Z + Wj) = \frac{W}{2}\dot{\omega} + j\frac{\dot{\omega}}{2}Z$$

The correction ΔY may be represented by a shift of the point $Y(G = 1\Phi = \pi)$ toward the point $Y + \Delta Y$ (Fig. 6 has been

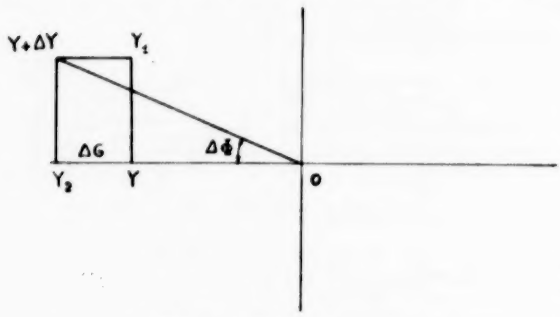


FIG. 6 PLOT OF INCREMENTS

drawn in the hypothesis of W and Z positive). The corrected gain is represented by the distance

$$OY_2 = 1 + \frac{W}{2}\dot{\omega}$$

where $\dot{\omega}$ is zero at the point Y, and $\dot{\omega}$ can thus be written for $\Delta\dot{\omega}$. Also, certain justified approximations are made in the figure such as the replacement of the distance from O to $Y + \Delta Y$ (gain) by the distance OY_2. The correction to be applied to the phase is represented by the angle $\Delta\Phi$, whose tangent is equal to

$$\frac{\overline{YY_1}}{OY} = \left|Z\frac{\dot{\omega}}{2}\right|$$

When $Z\dot{\omega}$ is positive, the phase angle becomes less than π, and the derivative of Φ is negative. Then we get the equation

$$\frac{1}{G}\frac{\partial G}{\partial\dot{\omega}} = \frac{W}{2}\frac{\partial\Phi}{\partial\dot{\omega}} = -\frac{Z}{2} \quad \ldots\ldots [19]$$

Value of the Approximation. The paper by Cahen and Loeb (7), summarized in the section, A First Approach, contains an attempt to predict the stability. (The second criterion is dealt with here; the first has been treated in the earlier section Criterion of Set of Curves.)

It was admitted, and we saw that this assumption is not rigorous, that the over-all transfer function G/Φ is analytic and represented by $U(x, \omega) + jV(x, \omega)$.

The second stability criterion was stated as follows: If the Jacobian determinant

$$J = \frac{\partial U}{\partial x}\frac{\partial V}{\partial\omega} - \frac{\partial V}{\partial x}\frac{\partial U}{\partial\omega}$$

is positive, the oscillation is stable. Here J may be written as

$$J = G\left[\frac{\partial G}{\partial x}\frac{\partial\Phi}{\partial\omega} - \frac{\partial G}{\partial\omega}\frac{\partial\Phi}{\partial x}\right] \quad \ldots\ldots [20]$$

In order to translate this result into the notation of Equation [14], we shall recall that the assumption of analyticity gives $\gamma - \beta = \pi/2$ and that in this crude theory the term d has been neglected. By inserting into Equation [20] the values drawn from Equation [17] it is seen that J has the same sign as

$$\sin(\alpha - \beta)\left(1 + x\frac{f'}{f}\right)$$

The second criterion gives the same result as the criterion derived from Equation [14] provided that $1 + xf'/f > 0$, and that d is neglected. The second criterion now appears to be valid for the limiting case of quasi-linear systems.

APPLICATION TO A PARTICULAR CASE—THE CONTACTOR SERVO

The contactor servo comprises a polarized relay whose transfer function may be calculated by the following procedure: The input applied to the coil is a sinusoidal voltage $x\cos\omega t$. It is assumed, as is actually the case, that the armature is attracted only when the input voltage is larger than a given value, that we call the "threshold" (let h be this value). The output is zero when $|x| < h$, and is equal to $\pm V_0$, where V_0 is the voltage of the battery, when $|x| > h$. The sign of the output is the same as the sign of the input. The output y is expanded in a Fourier series, and the first harmonic (fundamental) gives the transfer function $f(x)$ of the relay

$$\frac{y}{x} = f(x) = \left\{\begin{array}{c} 0 \text{ for } x < h \\ \dfrac{4V_0}{\pi x}\sqrt{1 - \dfrac{h^2}{x^2}} \text{ for } x \geq h \end{array}\right\} \quad \ldots\ldots [21]$$

Fig. 7 gives $f(x)$ plotted against x. The transfer function of the motor is

$$Y(s) = \frac{1}{Ts(1 + T_1 s + T_2^2 s^2)} = g/\varphi \quad \ldots\ldots [22]$$

Oscillation can take place for one value of ω, i.e., $1/T_2$. For this ω

$$Y\left(\frac{j}{T_2}\right) = -\frac{T_2^2}{TT_1}$$

and the phase of the output relative to the input is equal to π. We must then determine x in such a way that G becomes equal to 1. By algebraic manipulation the equation

$$G = \frac{T_2^2}{TT_1}\frac{4V_0}{\pi x}\sqrt{1 - \frac{h^2}{x^2}} = 1$$

becomes quadratic in x^2. By the well-known quadratic formula this equation has real solutions in x^2 when

$$\frac{TT_1}{T_2^2} < \frac{2V_0}{\pi h}$$

If this inequality is satisfied, the solutions for x^2 are positive and the corresponding values of x are real. In general, two are positive.

In order to apply the preceding results (Equation [10]) to the present case we must now calculate the two derivatives $\partial g/\partial\omega$ and $\partial\varphi/\partial\omega$. If $\eta = |Ts(1 + T_1 s + T_2^2 s^2)|$ with $s = j\omega$

$$\frac{1}{g}\frac{\partial g}{\partial\omega} = -\frac{\dfrac{\partial\eta}{\partial\omega}}{\eta}$$

At $\omega = 1/T_2$ the calculation gives

$$\frac{1}{g}\frac{\partial g}{\partial \omega} = -2T_2$$

In order to calculate $\partial\varphi/\partial\omega$ we must write the phase φ of the function

$$\frac{1}{Tj\omega(1 + T_1 j\omega - T_2^2\omega^2)}$$

One finds

$$\varphi = -\frac{\pi}{2} - \text{arc tan}\left(\frac{1 - T_2^2\omega^2}{\omega T_1}\right)$$

Then, for $\omega = 1/T_2$

$$\frac{\partial\varphi}{\partial\omega} = -\frac{2T_2^2}{T_1}$$

We will now calculate d and δ, with Equation [18]. As the transfer function of the linear part is analytic, the derivatives are calculated in the conventional manner

$$\left.\frac{d^2Y}{ds^2}\right]_{s=\frac{j}{T_2}} = Z + Wj = \frac{6T_1^2T_2^4 - 8T_2^6}{TT_1^3} - 10j\frac{T_2^5}{TT_1^2} \quad [23]$$

Then

$$\frac{1}{G}\frac{\partial G}{\partial\dot{\omega}} = -\frac{5T_2^5}{TT_1^2}\frac{\partial\phi}{\partial\dot{\omega}} = \frac{4T_2^6 - 3T_1^2T_2^4}{TT_1^3} \quad \ldots\ldots [24]$$

We then get all the needed coefficients

$$\left.\begin{aligned} a = f'/f,\ b\sin\beta &= -2T_2, \\ x_0c\sin\gamma &= -2(T_2^2/T_1)\left(1 + x_0\frac{f'}{f}\right) \\ \alpha = \pi/2,\ b\cos\beta &= -2\,T_2^2/T_1, \\ x_0c\cos\gamma &= 2T_2\left(1 + x_0\frac{f'}{f}\right) \\ d\sin\delta &= -\frac{5T_2^5}{TT_1^2}, \\ d\cos\delta &= \frac{4T_2^6 - 3T_1^2T_2^4}{TT_1^3} \end{aligned}\right\} \quad \ldots [25]$$

It happens quite often that T_1 and T_2 are of the same order of magnitude. Let us assume, in order to simplify the calculations, that $T_1 = T_2$. Then, referring to Equations [25], we get

$$\left.\begin{aligned} &a = f'/f,\quad b = 2T_1\sqrt{2},\quad x_0c = 2T_1\sqrt{2},\quad d = 5T_1^3/T \\ &\alpha = \pi/2,\quad \beta = \pi + \pi/4, \\ &\qquad\qquad \gamma = \pi/4 + 3\pi/2,\quad \tan\delta = -5 \end{aligned}\right\} \quad \ldots [26]$$

Equation [14] then becomes

$$-\frac{1}{K_1} = \frac{\dot{\xi}}{\xi} = \frac{\dfrac{2x_0f'}{f}}{8T_1\left(1 + x_0\dfrac{f'}{f}\right) + \dfrac{x_0f'}{f}\dfrac{T_1^2}{T}} \quad \ldots\ldots [27]$$

Let us draw from Equation [21] the value of the terms which depend on $f(x)$

$$\frac{xf'}{f} = \frac{2h^2 - x^2}{x^2 - h^2},\ 1 + \frac{xf'}{f} = \frac{h^2}{x^2 - h^2} \quad \ldots\ldots [28]$$

The correcting term

$$\left(1 + \frac{xf'}{f}\right)$$

is always positive, since x must be larger than h. We are now faced with two cases.

First Case. $f' > 0$ (point r_1 in Fig. 7): Here $1/K_1 = \dot{\xi}/\xi$ is positive. We already have seen in the general section that in this case the oscillation foreseen by theory cannot take place because it is unstable.

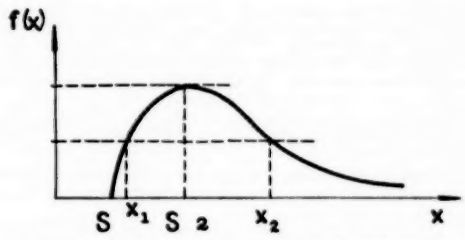


FIG. 7 TRANSFER FUNCTION OF RELAY WITH THRESHOLD

Second Case. $f' < 0$ (point x_2 in Fig. 7): Let us first assume explicitly that

$$\left|\frac{x_0f'}{f}\right|$$

is small compared to 1 and $8T/T_1$. That can happen if the abscissa x_2 of the second intersection in Fig. 7 is sufficiently near to $h\sqrt{2}$. The denominator will remain positive. It is then possible to foresee by this theory that the oscillating operation is stable.

In the general case, one has to carry out the analysis in more detail and in so far as we are aware, no general statement is possible. We know by experience that there certainly exists a stable operation. We know it by simple reasoning since an increase of x decreases the gain, and no physical amplitude can increase indefinitely. The derivation is not rigorous enough to be sure that no exception can occur.

CONCLUSION

Until recently, almost all the methods of calculation for nonlinear systems were strictly limited to the case of second-order equations.

Now it is well known that a servomechanism defined by an equation of second order cannot oscillate, unless the coefficient of the first derivative purposely has been made opposite in sign to the coefficient of the variable itself (in this case, the feedback is introduced in the wrong sense). The simplest really usable servo that gives rise to stability problems is then of the third order, and classical methods cannot be applied. The "filtered-systems" approach is then of great interest. Thanks to it, we are able to treat many problems correctly. For a wide class of nonlinear servos, we can predict self-sustained oscillations, and we are able to ascertain some cases where possible oscillations cannot build up.

ACKNOWLEDGMENTS

The author is greatly indebted to Mr. G. Cahen,[3] whose constructive criticism was of decisive importance in the building of the "Transient" theory. He also expresses deep gratitude to the Chairman of the ASME-IRD Dynamics System Committee, Dr. Rufus Oldenburger, who kindly invited him to con-

[3] Chief Engineer of the Navy (Naval Architect), Paris, France.

tribute to the present Symposium, and gave much helpful advice for the revision of this paper.

BIBLIOGRAPHY

1 "Théorie des Servomécanismes non Linéaires," by J. R. Dutilh, *La Radio Française*, May, 1950, pp. 1–7.

2 "Théorie des Servomécanismes à Relais," by J. R. Dutilh, *L'Onde Electrique*, vol. 30, 1950, pp. 438–445.

3 "Analyzing Contactor Servomechanisms by Frequency Response Methods," by R. J. Kochenburger, *Electrical Engineering*, vol. 69, August, 1950, pp. 687–692.

4 "Un Critérium Général de Stabilité des Servomécanismes, Sièges des Phénomènes Héréditaires," by J. M. Loeb, Comptes Rendus de l'Académie des Sciences, vol. 233, July 16, 1951, pp. 344–345.

5 "Phénomenes Héréditaires dans les Servomécanismes; un Critérium Général de Stabilité," by J. M. Loeb, *Annales des Télécommunications*, vol. 6, December, 1951, pp. 346–352.

6 "Perturbations des Oscillateurs Filtrés," by G. Cahen, Comptes Rendus de l'Académie des Sciences, vol. 235, December 22, 1952, pp. 1614–1617.

7 "Calcul de l'Amortissement dans les Oscillateurs Filtrés," by G. Cahen and J. M. Loeb, *Annales des Télécommunications*, vol. 8, March, 1953, pp. 97–101.

8 "Variable Frequency Electric Circuit Theory With Application to the Theory of Frequency Modulation," by J. R. Carson and T. C. Fry, *Bell System Technical Journal*, vol. 16, October, 1937, pp. 513–540.

Discussion

R. J. Kochenburger.[4] The author should be commended for the work he has done in advancing the theory of nonlinear servomechanisms as represented by this and his previous papers. It is believed that these contributions have provided additional mathematical justification to those techniques where the frequency-response approach is applied to such nonlinear systems. His work also has pointed out carefully the limitations imposed on this approach.

The writer wishes to supplement the paper by mentioning several other contributions in this field. The original suggestion regarding the concept of neglecting higher harmonics first arose in the textbook by Oldenbourg and Sartorius (9).[5] The author has mentioned contributions made by J. R. Dutilh and this discusser; it might also be mentioned that Goldfarb (10) in the USSR, arrived at similar conclusions at about the same time. Goldfarb's work, however, was not brought to the general attention of engineers in the U. S. and Western Europe until some number of years later. Following these original contributions, E. C. Johnson (11) extended this approach in a number of respects. Among other suggestions and conclusions, he proposed a means for including the effects of some of the more significant higher-harmonic components.

Since that time other papers have appeared which apply this general "describing function" method to specific types of servomechanism problems. One by V. B. Haas (12) applies this method to systems which include significant friction effects. One by the discusser (13) applies the method to instances where limiting and saturation effects are important. This latter, and a paper by E. S. Sherrard (14) discuss the difficulties that can arise when limiting effects appear in conditionally stable systems and suggest means for eliminating these difficulties.

It is believed that many further advances employing this general technique can be expected to advance the theory of nonlinear control-system design in the near future. The author, for example, calls attention to a number of problems that still await solution.

[4] Professor of Electrical Engineering, University of Connecticut, Storrs, Conn.

[5] Numbers in parentheses apply to the Bibliography references (9 to 14, inclusive) at the end of this discussion.

BIBLIOGRAPHY

9 "Dynamics of Automatic Control," by R. C. Oldenbourg and H. Sartorious, translated by H. L. Mason, ASME, New York, N. Y., 1948.

10 "On Some Non-Linear Phenomena in Regulatory Systems," by L. C. Goldfarb, *Avtomatika i Telemekhanika* (Russian), vol. 8, no. 5, 1947, pp. 349–383; translated in National Bureau of Standards Report No. 1691, 1952.

11 "Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements," by E. C. Johnson, Trans. AIEE, vol. 71, 1952, Part II, *Applications and Industry*, pp. 169–181.

12 "Coulomb Friction in Feedback-Control Systems," by V. B. Haas, Jr., Trans. AIEE, vol. 72, 1953, Part II, *Applications and Industry*, pp. 119–126.

13 "Limiting in Feedback Control Systems," by R. J. Kochenburger, Trans. AIEE, vol. 72, 1953, Part II, *Applications and Industry*, pp. 180–194.

14 "Stabilization of a Servomechanism Subject to Large Amplitude Oscillation," by Emile S. Sherrard, Trans. AIEE, vol. 71, 1952, Part II, *Applications and Industry*, pp. 312–324.

W. H. Surber, Jr.[6] This paper outlines some aspects of a method of generalizing the "frequency-response" methods of analyzing linear systems so that they may be applied approximately to nonlinear systems. If valid, this method may be used to predict the stability of the system and to locate approximately the phase space trajectories or "limit cycles" of unstable systems. This is an attractive possibility, of course, due to the well-known difficulties in handling nonlinear differential equations of the third order or higher, and it has been studied by a number of investigators.

The method assumes that the linear part of the system completely filters out the harmonic wave-form distortion produced by the nonlinear element, i.e., that these harmonics are not transmitted around the loop and do not appear at the input to the nonlinear device. This is justified approximately in many cases. Fortunately, the more complex the system, the better the method usually applies. Since it is potentially so useful, however, it would be desirable to place it on a more rigorous foundation if possible. One paper, by E. C. Johnson (11), has appeared in which a method taking these nonlinearly generated harmonics of the signal wave form into account is described. Although the calculations are rather complex, it provides a valuable method of checking the applicability of the more approximate approach.

R. H. Macmillan.[7] What is now generally known as the "describing function" method, as first applied by Tustin to hereditary nonlinearities in linear servosystems and by Kochenburger to relay-control systems, involves the plotting of a harmonic response locus $Y(\omega)$ for the linear elements of the servo loop, together with a second locus, or describing function, $-1/f(x)$, which is the reciprocal of the amplitude-dependent gain of the nonlinear elements. Intersections of these loci may represent the possibility of sustained oscillation at the amplitude and frequency indicated.

When the amplitude and frequency-dependent elements are separable, it is rather easy to see from a sketch whether an intersection of the loci represents a stable or an unstable mode of oscillation. Consider, for example, the loci indicated in Fig. 8, herewith, intersecting at *A*, *B*, and *C*.

If the locus $Y(-\alpha + j\omega)$ is plotted, as shown by the dotted line, it will lie close to $Y(j\omega)$ for small values of α and on the side of that locus indicated. Intersections of this curve with the describing function locus represent states of positively damped oscillation; i.e., the amplitude tends to decrease.

Suppose now that oscillation is occurring at the point *A* when its amplitude is slightly increased, so moving the operating point

[6] Associate Professor of Electrical Engineering, Princeton University, Princeton, N. J.

[7] University Lecturer, Cambridge University, Cambridge, England.

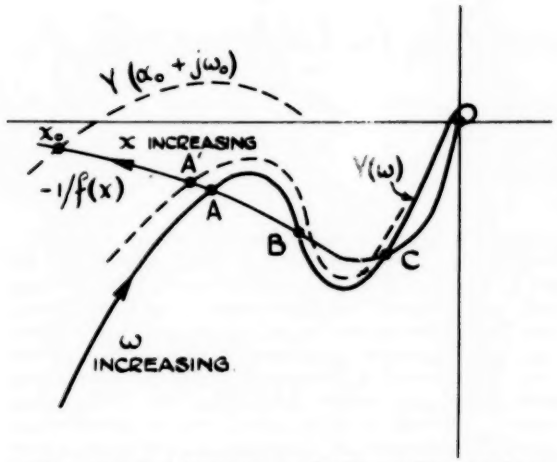


FIG. 8 HARMONIC RESPONSE AND DESCRIBING FUNCTION LOCI

along the describing function locus to A'. We have shown that the resulting motion at this point is one of decreasing amplitude, so that the oscillation tends to return to its equilibrium value at A, which is thus a stable point. One can show similarly that B is an unstable point and C is stable. This alternation of stable and unstable limit cycles is in accordance with the general theory propounded by Poincaré.

It will be noted that if the system is started oscillating freely with amplitude x_0, it would appear possible to determine the nature of the initial mode of damped oscillation, $x_0 e^{-\alpha t} \cos \omega_0 t$, from the curve $Y(\alpha_0 + j\omega_0)$ on which the point x_0 of the describing function lies. Such an elementary approach ignores the rate of change of damping and frequency and can lead to appreciable error. When $\dot{\alpha}$ and $\dot{\omega}$ are taken into account, however, relatively simple analytical expressions can be derived for the transient response of nonlinear systems; these are as accurate as the results obtainable by the describing function method for the sustained oscillation. This work was done by Mr. P. E. W. Grensted at Cambridge University and will be published shortly.

When the amplitude and frequency-dependent elements are not separable, it would be possible to regard the describing functions as a series of contours for each frequency, the whole being plotted so as to form a surface with frequency as a third axis normal to both those already used. This surface would have upon it lines of constant x. The $Y(\alpha + j\omega)$ lines form a second surface with lines of constant α. These two surfaces intersect, in general, in an "operating line," to every point of which there corresponds a single value of both x and α. Points of possible continuous oscillation are those where $\alpha = 0$. A small disturbance at such a point would move the system along the operating line and the sign of α determines, as formerly, the state of equilibrium of the operating conditions considered. This is a graphical interpretation of what M. Loeb has done analytically in his paper; analysis is indispensable for any nonseparable system, except the very simplest, but the foregoing approach may help toward a physical understanding of the equations.

R. L. COSGRIFF.[8] The author is to be congratulated on his excellent paper. The writer was particularly interested in his discussion of methods for determining the stability of limit cycles, having used the method for determining the stability of several systems that the writer has studied and found that the results so obtained were reliable in all cases except one. The exception occurred in a system whose open-loop frequency response exhibited jump phenomenon or response hysteresis. The difficulty arises from the fact that the gain curves are discontinuous and may not pass through the —1 point which is point N, Fig. 4 (a) of the paper.

The writer agrees with the author's conclusion that the frequency-response approach can be used in the analysis of many nonlinear systems which cannot be handled by other methods. He would like to call attention to the fact that the frequency-response approach for nonlinear systems only indicates the system stability, while other performance criteria based upon the frequency response of linear systems are not always reliable when used for nonlinear systems.

AUTHOR'S CLOSURE

I am grateful to the discussers for the great amount of information they contributed to the subject of the paper. First among the discussers to be acknowledged is Prof. R. J. Kochenburger who gave a careful survey of literature that is not familiar in our country.

I agree fully with Prof. W. H. Surber's statement: "Fortunately, the more complex the system, the better the method usually applies." That is, the more involved are the high-degree transfer functions, the more accurate becomes the filtering assumption.

Dr. R. H. Macmillan points to an equivalence between my limit-cycle stability criterion and his graphic example in Fig. 8. However, it is established with the tacit assumption that the describing function is analytic. We saw that this is not so, and that the analyticity assumption can be considered as a limit assumption.

I am in agreement with Prof. R. L. Cosgriff's statements. Up to the present, the "describing function" method has been able to solve stability problems but gives very little information about other requirements such as response speed.

I would like to mention that until now one point has escaped the attention of various writers; namely, the equations that are presented in the paper are true under the assumption that the nonlinear part gives an output comprising only the fundamental and upper harmonics. But it also can give a d-c component that is not filtered by the linear part of the servomechanism. A paper concerning this subject will soon be published by the *Annales des Télécommunications* under the signature of J. D. Lebel (Schlumberger Well Surveying Company) and the author.

[8] Assistant Professor, Department of Electrical Engineering, Ohio State University, Columbus, Ohio.

A Statistical Approach to Servomechanisms and Regulators

By M. J. PÉLEGRIN,[1] PARIS, FRANCE

The servo field is developing rapidly in all of its aspects. The methods presently used are based upon the theory of analytic functions and consist mainly of the study of the system subjected to steady sine-wave disturbances. Recently it has been proved that, in certain cases, it is possible to compute the system directly from statistical data concerning inputs (inputs composed of "orders" and "disturbances"). The paper first reviews the results that were obtained recently by the author in this field. A particular servocircuit is computed by statistical methods, and the output autocorrelation function of a relay also is obtained. Classical and statistical methods are then used in the analysis of a variable-inertia regulator. The principal properties of the variable-inertia regulator are derived for the cases where the inertia is controlled through an external source of power and where the inertia is self-adjusted by a passive network.

Introduction

GENERALLY, servomechanisms are studied for the case of steady oscillations, i.e., a periodic steady state. If the system is a linear one, i.e., if the equations which summarize it are linear, the superposition principle is valid and we can consider the response of the system to a periodic input function as the summation of the individual response to the various harmonics of the input function. (In this paper we call the input function either "order" or "message"; however, the input function can be the sum of a message and a noise.) Generally, the coefficients of the differential-equation system can be treated as constant—for example, an electrical system with lumped coefficients, or a mechanical linkage in which the mass of the springs can be neglected. The response to a periodic disturbance is also periodic and can be characterized by the amplitude ratio composed of the output divided by the input, and the phase lag of the output relative to the input. We then obtain the amplitude versus frequency, and phase versus frequency. These two diagrams can be summarized in one curve on the complex plan. This is the transfer locus of the system. As soon as the system is no longer linear with constant coefficients these diagrams do not fit and the amplitude-phase concept, which exists for linear systems disappears (1).[2]

Linear systems can represent many nonlinear systems, as a first approximation. We can then deduce from the associated diagrams the response to a specific input. But this work is long and superficial because the inputs to which the system will have to respond are generally unknown a priori. It is possible to conceive of the association of a "probability of occurrence" with certain inputs and to study in full detail the response to the more probable ones.

This paper goes a step further. In many cases the input is a random time function (we shall later define the exact meaning of the mathematical terms used) in which case the system will have to respond to statistically equivalent functions. However, two records of the input made at different times will not be identical but the statistical analysis of these records will give the same results. If we obtain the same statistical results whatever the initial instant, the input function is said to be "stationary." We shall consider only such functions. We must note that a system is always fed with noise, which, by definition, is represented by a random function.

From such statistical data it is possible to compute a system so that it will be optimum for given criteria. A criterion which leads to relatively simple mathematical results is the "mean-square error." The error is computed from the difference between the actual output and the "desired" output function; it is understood that the input function is a random function statistically well defined. The problem consists of making that error a minimum by determining a corrective device (electrical network, mechanical linkage, and so on). Some definitions follow (2, 3).

Random Function. Function which cannot be known exactly at any time t, but of which each "sample" has the same statistical properties, namely, mean value, quadratic mean value, repartition function and autocorrelation function.

Sample of a random function is a known function of time which is obtained by taking a record of the random function during t seconds, beginning at the instant T. Samples of a random function cannot be superimposed.

Probability Distribution Function of a random function: If x is the independent variable, this is the function $F(X)$ where $F(X)$ denotes the probability that x be less than or equal to X. The derivative of the repartition function $F(X)$ is the probability $p(x)$ that x occurs between X and $X + dX$

$$p(x) = \left[\frac{dF(X)}{dX}\right]_{\text{for } X = x} \qquad [1]$$

Characteristic Function. This is the Fourier transform of probability distribution function, or the mean value of e^{iux} given by

$$\phi(u) = \int_{-\infty}^{+\infty} e^{iuX} dF(X) \qquad [2]$$

The expansion in powers of u can be written as

$$\phi(u) = \sum_m \frac{(iu)^m}{m!} M_m \qquad [3]$$

where M_m is the mth "moment" of x

$$M_m = \int_{-\infty}^{+\infty} x^m p(x) dx \qquad [4]$$

and the summation in Equation [3] is over $m = 0, 1, 2, \ldots$.

---

[1] Engineer, Service Technique Aeronautique.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 26, 1953. Paper No. 53—A-20.

Mean Value. Two mean values must be envisaged. If $f(t)$ is a (known or random[3]) function its algebraic-mean value $\overline{f(t)}$ is given by

$$\overline{f(t)} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{+T} f(t)dt \quad \ldots\ldots\ldots\ldots [5]$$

If x is a random variable which takes the value x with the probability $p(x)$, its probability-mean value (mathematical expectancy) is $\bar{x}$ where

$$\bar{x} = \int_a^b x\, p(x)\, dx \quad \ldots\ldots\ldots\ldots [6]$$

By virtue of Equation [1] this also can be written as

$$\bar{x} = \int_a^b x\, dF(x) \quad \ldots\ldots\ldots\ldots [7]$$

where a and b are the lower and upper limits of x.

Auto and Cross-Correlation Function of a random or known function $f(t)$. This is the function of τ which is the mean value $\varphi(\tau)$ of the product of $f(t) \times f(t + \tau)$ where τ is a parameter, i.e.

$$\varphi(\tau) \triangleq \overline{f(t)f(t + \tau)} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{+T} f(t)f(t + \tau)\, dt \quad [8]$$

Because of the limitation of the study to stationary random functions we can make use of the ergodic theorem[4] (5) and give another expression for the autocorrelation function.

Suppose we can get a great number of samples of the random function, either by making a great number of recordings during τ seconds (τ is a parameter which is kept constant for a number of samples) or by cutting off a long record of the random function in samples of τ seconds duration each, Fig. 1.

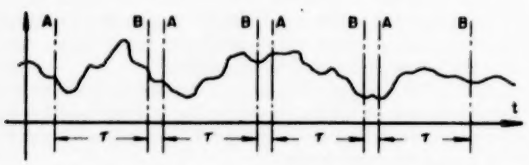


FIG. 1 SAMPLES (OF τ SECONDS DURATION) OBTAINED FROM A RECORD OF THE RANDOM FUNCTION
(Note that samples can be adjacent.)

Let us call A the vertical axis which is at the origin of each sample and B the one which is at the end (A and B are separated by τ seconds, Fig. 2). Term a_i denotes the ith segment (supposed to be very small) on axis A, b_j the jth segment on axis B. From each set[5] of samples having the same duration, we can compute the following:

The "probability" $p(a_i)$ that the curve cut the first axis A within the ith segment a_i.

The "conditional probability" $p(b_j/a_i, \tau)$ that the second axis B be cut within b_j when the first axis is cut within a_i.

The probability-mean value of the product of $f(t)$ by $f(t + \tau)$ can be written as

$$\varphi(\tau) = \sum_{i/j} a_i p(a_i) b_j p(b_j/a_i, \tau) \quad \ldots\ldots\ldots\ldots [9]$$

[3] We recall that a random function should be denoted as $f(t, E)$ where E is a "parameter" which can be used to define the samples. For details see reference (4).

[4] This theorem states that for stationary random functions algebraic and probability means are equal.

[5] Each set is supposed to be composed of a great number of samples.

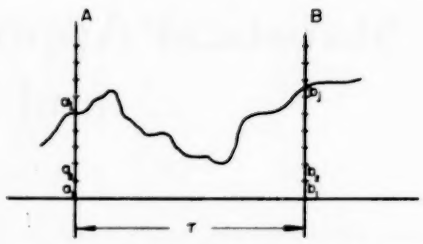


FIG. 2 ENLARGEMENT OF SAMPLE IN FIG. 1
(Axes A and B are divided into small segments.)

If $f_1(t)$ and $f_2(t)$ are two random (or known) functions, the "cross-correlation" function is

$$\varphi_{12} \triangleq \overline{f_1(t)f_2(t + \tau)} \quad \ldots\ldots\ldots\ldots [10]$$

Let

$$f_T(t) = \begin{cases} f(t) & \text{for } 0 < t < T \\ 0 & \text{elsewhere} \end{cases} \quad \ldots\ldots\ldots\ldots [11]$$

The corresponding Fourier integral is

$$F_T(\omega) = \int_{-T}^{+T} f_T(t)\, e^{-j\omega t}\, dt \quad \ldots\ldots\ldots\ldots [12]$$

The "energy spectrum" of $f(t)$ is by definition represented by

$$\psi(\omega) = 2\pi |F_T(\omega)|^2 \quad \ldots\ldots\ldots\ldots [13]$$

and the power density spectrum $\phi(\omega)$ is

$$\phi(\omega) = \lim_{T \to \infty} \frac{\pi}{T} |F_T(\omega)|^2 \quad \ldots\ldots\ldots\ldots [14]$$

Kintchine pointed out that the two functions $\varphi(\tau)$ and $\phi(\omega)$ are a pair of Fourier transforms, given by

$$\varphi(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \phi(\omega)\, e^{j\omega\tau}\, d\omega \quad \ldots\ldots\ldots\ldots [15]$$

$$\phi(\omega) = \int_{-\infty}^{+\infty} \varphi(\tau) e^{-j\omega\tau}\, d\tau \quad \ldots\ldots\ldots\ldots [16]$$

GENERAL RESULTS CONCERNING LINEAR SERVOMECHANISMS

The starting point of the theory that we have developed (6) deals with the theory of the "optimum filter" by Prof. Norbert Wiener and Prof. Y. W. Lee at the Massachusetts Institute of Technology. The problem treated by Professor Lee (2) is the following: A known message is to be sent. It is supposed to be a random stationary time function mixed with a noise which also is a random stationary function. The transfer function $H(s)$ of the filter is computed in such a way that the "mean-square error" between the actual output and the desired output is a minimum. It has been proved that only the autocorrelation function of the input and the cross-correlation function between the input and the desired output are necessary to solve the problem (4, 2)—(Wiener-Hopf equation). Professor Lee has developed a satisfactory method of solving it (2).

We now consider a servomechanism, the input of which is a stationary random function. The servo also is disturbed by a noise (stationary random function) which enters the loop at the "torque point,"[6] Fig. 3.

[6] This is a first difference with the optimum filter theory. If noise is introduced only at the very input of the servo, the Wiener-Lee theory can easily be adapted to the servo field.

The first block represents the transfer function of the corrective network and is the unknown in the problem. The second block represents the mechanical component and is generally assigned. Thus $B(s)$ may be the transfer function of a radar antenna with inertia and damping. The variable θ_p may be associated with disturbances due, for example, to wind gusts. Moreover, we can assume that noise also is superimposed at the input to the system, but for the sake of simplicity of computations we shall assume that θ_e is a message without noise.

We break down the block diagram of Fig. 3 into the one in Fig. 4 which is equivalent and no longer looped. We apply to each branch a method similar to the Wiener-Lee method and we can go straight to the result, but computations are very lengthy.

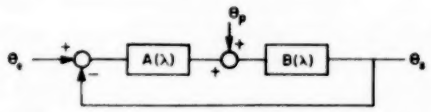


FIG. 3 CONVENTIONAL SERVO WITH TWO INPUTS

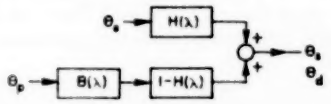


FIG. 4 EQUIVALENT BLOCK DIAGRAM: $H(s)$ REPRESENTS $\dfrac{AB}{1+AB}$

However, in the case in which the desired output function θ_d is identical to the input function (which has been assumed to be without noise) computations are very simple and can be summarized here.

By the calculus of variations and a Fourier transformation[7] the expression for the mean-square error leads to the following condition:

We introduce $P(s)$, where

$$P(s) = \phi_{ee}(s) + |B(s)|^2\phi_{pp}(s) \quad [17]$$

$\phi_{ee}(s)$ and $\phi_{pp}(s)$ are the power spectra of the input and disturbing functions, respectively, $B(s)$ is the given transfer function, $H(s)$ is still the unknown transfer function. We introduce $Q(s)$ where

$$Q(s) = P(s)\,[H(s) - 1] \quad [18]$$

The condition is that the poles of $Q(s)$ must be located in the right half plane only.

This condition comes from the physical condition $h(t) \equiv 0$ for $t < 0$, where $h(t)$ is the impulse response of $H(s)$. By means of the operations that define the auxiliary function $Q(s)$ this physical condition yields the foregoing condition on the "poles" of $Q(s)$.

If we restrict ourselves to rational solutions, the solution of this problem for $H(s)$ is unique (6).[8]

Equation [17] shows that $P(s)$ can be assumed to be rational, because spectra which come from autocorrelation functions can be approximated by rational expressions.

We now reach the following solution:

(a) After having broken down both numerator and denominator of $P(s)$ into linear factors, we must build $H(s) - 1$ in such a way that its "zeros" are equal to the poles of $P(s)$ which are located in the left half plane, in order to fulfill the foregoing condition, i.e., no poles in the left half plane for $Q(s)$.

(b) An expression for $H(s) - 1$ consisting only of linear factors in the numerator and nothing in the denominator is unacceptable. The function $H(s)$, as any transfer function of a real component, must tend toward 0 when s tends toward infinity. It is then necessary to put a number of poles for $H(s) - 1$ equal to the number of zeros already determined.

(c) For stability reasons the transfer function $H(s)$ can have poles only in the left half plane. The poles of $H(s) - 1$ are the same as those of $H(s)$. They should be located in the left half plane. They consequently can be taken among the zeros of $P(s)$ which are located in the left half plane in order that the foregoing condition concerning the poles of the function $Q(s)$ be fulfilled.

(d) The number of such zeros may not be equal to the number of poles already determined. In that case (we have proved that there can only be the difference of one linear factor) we introduce an arbitrary linear factor either in the denominator or numerator of $H(s) - 1$ and we determine the two constants involved in that linear factor by the fact that:

1 $H(s) - 1$ must tend toward -1 when $|s|$ tends toward infinity.

2 And $H(s)$ must tend toward zero at least as $1/r^2$ where r is the modulus of s. (Theory of analytic functions: $h(0) = \int_{-\infty}^{+\infty} H(s)ds$ must be equal to zero because the impulse response $h(t)$ is equal to zero when $t = 0$.)

The following numerical example will show the procedure. We consider a servo represented by Fig. 3 in which the input is the stationary function shown in Fig. 5 and statistically defined by a Poisson distribution for the instants of crossing the t-axis. The probability $p(n, t)$ of having n crossings within time t is given by

$$p(n, t) = e^{-kt}\frac{(kt)^n}{n!} \quad [19]$$

where k is the average number of crossings per unit of time.

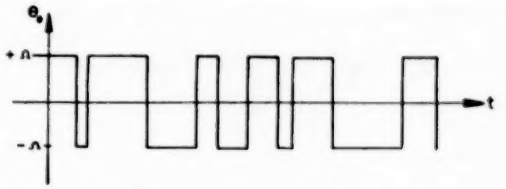


FIG. 5 STATIONARY RANDOM FUNCTION $\theta_e(t)$

We have chosen numerical values so that the calculations are simple. The spectrum of the random function shown in Fig. 5 is

$$\phi_{ee}(s) = \frac{\Omega^2}{\pi}\frac{2k}{(2k)^2 - s^2} = \frac{\sqrt{\frac{36}{13}}}{\frac{36}{13} - s^2} \quad [20]$$

We choose for the disturbances a flat spectrum given by

$$\phi_{pp}(s) = \sqrt{36 \times 13} \quad [21]$$

Finally, we assume that $B(s)$ represents an inertia without damping

$$B(s) = \frac{1}{s^2} \quad [22]$$

[7] Expression [14] is similar to a Fourier transform of a Wiener-Hopf equation (see reference 6).

[8] All the results mentioned here are proved in reference (6).

Expression [18] denoted $Q(s)$ can be written as

$$Q(s) = \sqrt{\frac{36}{13}}\,\frac{s^4 - 13s^2 + 36}{s^4\left(\frac{36}{13} - s^2\right)}\,[H(s) - 1] \quad \ldots\ldots [23]$$

that is

$$Q(s) = \sqrt{\frac{36}{13}}\,\frac{(s+2)(s-2)(s+3)(s-3)}{s^4\left(s + \sqrt{\frac{36}{13}}\right)\left(s - \sqrt{\frac{36}{13}}\right)}\,[H(s) - 1] \quad [24]$$

from which we "build"

$$H(s) - 1 = K\,\frac{\left(s + \sqrt{\frac{36}{13}}\right)(s+u)}{(s+2)(s+3)} \quad \ldots\ldots [25]$$

where K is a constant and $s + u$ is a linear factor, both to be determined. We see immediately that $K = -1$ because $H(s)$ must tend toward zero when $|s|$ tends toward infinity, and we then deduce

$$H(s) = \frac{(s+2)(s+3) - \left(s + \sqrt{\frac{36}{13}}\right)(s+u)}{(s+2)(s+3)} \quad \ldots [26]$$

To have $|H(s)|$ decrease at least as $1/r^2$ ($r = |s|$) we have to cancel the highest-power term in the numerator. This finally leads to the expression

$$H(s) = \frac{6 - \sqrt{\frac{36}{13}}\left(5 - \sqrt{\frac{36}{13}}\right)}{(s+2)(s+3)} = \frac{0.45}{(s+2)(s+3)} \quad [27]$$

We have obtained the same value by the Wiener-Lee method; computations, in that general case, are very lengthy.

Statistical Study of Relays

A relay, essentially composed of a coil, a magnetic circuit, and an armature, is a nonlinear system the output of which can be represented either by $+1$ or 0 according to the position of the contact (closed or open). We study the problem of the determination of the output spectrum, knowing the statistical properties of the random function that represents the voltage input. When the two dead times of the relay (associated with closed-open and open-closed) are identical it is easy to determine the output spectrum. Actually, the two dead times are very different, and this is the reason which justifies the study of this problem.

The first case to be envisaged concerns a relay which is energized by stationary random voltages, much greater than the threshold of the relay. Only the instants of crossing the t-axis are of interest (sign relay), Fig. 6. Let us call $p(n, t)$ the probability of having n crossings within time t; that function is known either by computation from the data available on the input function (see the example in the preceding section) or by computation from records (samples) of the stationary input function. Let $p(+)$ be the probability that the input voltage be positive and $p(-)$ the opposite. We derive the output autocorrelation function of the relay from these data (6). This function is

$$\varphi(\tau) = p(+)\int_{\theta_1}^{\infty} p(0, t)dt\left[p(0, \tau - \theta_1) + \sum_{\text{even}}^{n} p(n, \tau - \theta_2) + (\theta_2 - \theta_1)\frac{\partial p(1, \tau)}{\partial \tau}\right] + p(-)\int_0^{\theta_1} p(1, t)dt$$

$$\left[\left(\frac{\theta_1 - \tau}{\theta_1}\right)_{\tau > \theta_1} + \sum_{\text{odd}}^{n} p(n, \tau - \theta_2) + (\theta_2 - \theta_1)\frac{\partial p(1, \tau)}{\partial \tau}\right] \quad \ldots\ldots [28]$$

where θ_1 and θ_2 denote the two time constants (opening and closure).

We obtain similar results in the case where the relay is subjected to varying voltages very close to the threshold of the relay (balance relay).

In addition to that theory we measured the departures of the dead times θ_1 or θ_2 which can be very important when the relay is attached to a vibrating support (frame of a piston-engine airplane). In the space available here it is not possible to comment on the experiment and the mathematical derivation which leads to the results (for details see reference 6).

For general nonlinear systems we have not yet obtained significant results.

Study of a Variable-Inertia Regulator

We shall try to indicate how to use this theory for the computation of a speed regulator composed of a variable-inertia regulator. We state the problem as follows:

A main motor (three-phase motor, for example) is supposed to drive a shaft at a constant speed regardless of the output load. We assume only that the fluctuations of the required power drawn from the electric motor can be represented by a stationary random function (with a mean value which generally is not zero but corresponds to the mean power required).

In order to avoid the employment of an underloaded electric motor, we can imagine the use of a device which will furnish or absorb power according to the fluctuations only (the average power will be furnished by the electric motor). We suggest the use of a Watt governor in which the position of the two balls is controlled by an external source of power or is self-adjusted by a passive system (7), Fig. 7.

The balls, being at a distance r_1 from the shaft axis, act as a flywheel as long as that distance does not change; but if at another time they are at a distance r_2 from the shaft, the kinetic

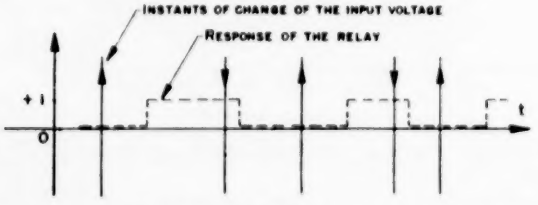


Fig. 6 Response of a Relay or Sign Relay, to a Random Input Voltage

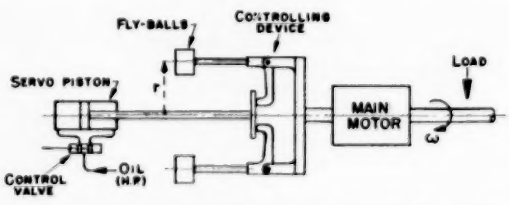


Fig. 7 Controlled Variable-Inertia Regulator With Hydraulic Controlling Device

energy will go from $mr_1^2\omega^2$ to $mr_2^2\omega^2$ where ω is the angular speed of the shaft (supposed to be kept constant) and m is the mass of each ball.

The variation of energy appears during the radial motion of the balls. The system is a source of energy if the balls move toward each other, an energy absorber in the opposite case. We use the following notations:

m = total mass of balls
r = distance between balls and shaft (variable)
$\omega = \dfrac{d\theta}{dt}$ = angular speed of shaft
I_1 = inertia of system without balls
C = opposing torque (constant), negative
$C(t)$ = opposing torque, function of time
$c(t) = C(t) - \overline{C(t)}$ (see footnote[9])
$\eta = 1/\mu$ = efficiency of control device

We will study the case in which no electric motor is connected to the shaft.

We start from the fundamental equations of the motion of a point in polar co-ordinates (r, θ)

$$\Gamma_T = r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt} \quad \ldots\ldots [29]$$

$$\Gamma_R = \frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2 \quad \ldots\ldots [30]$$

where Γ_T and Γ_R are the components of the acceleration of the point having polar angles respectively equal to θ and $\theta + \pi/2$.

According to these notations, we shall use a more usual form of Equation [29] which is

$$(I_1 + mr^2)\frac{d\omega}{dt} + 2m\omega r\frac{dr}{dt} = C \quad \ldots\ldots [31]$$

where C is the torque with regard to the origin of co-ordinates.

Multiplying both sides by ωdt we get

$$(I_1 + mr^2)\,\omega d\omega + 2m\omega^2 r dr = C\omega dt \quad \ldots\ldots [32]$$

The equation now concerns energy quantities.

To move the balls a distance of dr, work

$$dE = m\omega^2 r dr \quad \ldots\ldots [33]$$

is involved. It is furnished to the system if dr is negative (centrifugal force).

Let us assume that ω is kept constant when r varies by dr. If dr is negative a brake (or an equivalent device) should be used in order to dissipate energy. This energy to be dissipated is given by Equation [32] in which $d\omega = 0$. This is

$$2m\omega^2 r dr = C\omega dt \quad \ldots\ldots [34]$$

The comparison between Equations [33] and [34] shows that the system delivers (or absorbs) an energy $2dE$ and to do so requires an energy dE. We note that the kinetic energy is decreasing and the process will stop when r reaches zero.

We assume that the initial speed of the system, which will now be called a regulator, is ω_0 at $t = 0$ and that $r = R$. We assume that the energy to move the balls is drawn from the regulator itself (efficiency η). The equation defining the motion is

$$(I_1 + mr^2)\frac{d\omega}{dt} + 2mr\omega\frac{dr}{dt} = C + \mu m\omega r\frac{dr}{dt} \quad \ldots\ldots [35]$$

[9] $\overline{C(t)}$ means: "mean value of $C(t)$."

The second term of the second member represents the torque required to move the balls.

If we are looking for a constant speed we must have

$$m\omega r\frac{dr}{dt} = \frac{C}{2-\mu} \quad \ldots\ldots [36]$$

Equation [36] defines the motion of the balls.

The integral satisfying the initial conditions is

$$r^2 = R^2 + \frac{2Ct}{m\omega_0(2-\mu)} \quad \ldots\ldots [37]$$

The speed regulation vanishes after a time t_1 given by

$$t_1 = \frac{m\omega_0 R^2}{C}\left(\frac{\mu}{2} - 1\right) \quad \ldots\ldots [38]$$

The efficiency of the controlling device must be greater than 0.5.

If we had used a constant-inertia wheel with inertia equal to the maximum of the last one ($I_1 + mR^2$), the speed would have decreased linearly and would have become equal to zero for t_2 given by (assuming that C is negative)

$$-t_2 = \frac{I_1\omega_1}{C} + \frac{m\omega_1 R^2}{C} \quad \ldots\ldots [39]$$

Fig. 8 shows the speed versus time for the two systems (unbroken curve corresponds to $\eta = 0.8$).

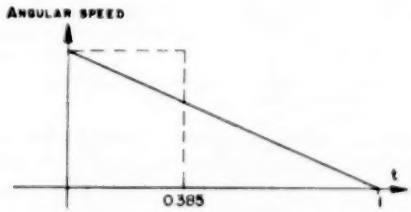


FIG. 8. DIAGRAMS OF SPEED VERSUS TIME FOR CONSTANT-INERTIA WHEEL—UNBROKEN LINE; AND VARIABLE-INERTIA WHEEL—BROKEN LINE
(Efficiency of mechanism to move balls, $\eta = 0.8$.)

The comparison of slopes at $t = 0$ for the two wheels when a particular formula for dr/dt is imposed is of special interest. Let us assume that the simplest law is studied

$$\frac{dr}{dt} = -A \text{ (const)} \quad \ldots\ldots [40]$$

This assumption applied to Equation [31] or [35] (in which $\mu = 1$), to which we add $I_1 = 0$, leads to the following equation which includes the two cases according to the value of α

$$m(R - At)^2\frac{d\omega}{dt} - \alpha Am\omega(R - At) = C \quad \ldots\ldots [41]$$

$\alpha = 1$ corresponds to the case in which the energy for moving the balls is taken from the regulator itself (with $\eta = 1$; otherwise $\alpha < 1$).

$\alpha = 2$ corresponds to the case where that energy is supplied to the system by an external source.

The integral of Equation [41] which satisfies the condition $\omega = \omega_0$ for $t = 0$ is

$$\omega = \frac{C}{mA(R - At)}\log\frac{R}{R - At} + \frac{R\omega_0}{R - At} \quad \text{if } \alpha = 1 \quad [42]$$

$$\omega = \left(\frac{C}{m}\,t + \omega_0 R^2\right)\frac{1}{(R - At)^2} \quad \text{if } \alpha = 2 \qquad [43]$$

The slope is (note that $C < 0$)

$$\left(\frac{d\omega}{dt}\right)_R = \frac{C + \alpha A m \omega\,(R - At)}{m\,(R - At)^2} \qquad [44]$$

while the constant inertia flywheel leads to the equation

$$\left(\frac{d\omega}{dt}\right)_F = \frac{C}{mR^2} \qquad [45]$$

We shall compare the variations of speed, for $C < 0$ near $t = 0$, where t is small compared to R/A. Now

$$\frac{\left(\frac{d\omega}{dt}\right)_R}{\left(\frac{d\omega}{dt}\right)_F} = \frac{1}{\left(1 - \frac{A}{R}\,t\right)^2} + \alpha\,\frac{A\omega mR}{|C|}\,\frac{1}{\frac{A}{R}\,t - 1} \qquad [46]$$

See Fig. 9 in which N is a parameter defined by

$$N = \alpha\,\frac{A\omega mR}{|C|} \qquad [47]$$

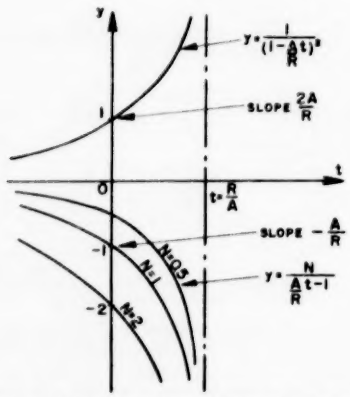


Fig. 9 Graph for Speed-Control Problem

We note that for $N = 1$ and t near 0 the second member of Equation [46] is near zero. We then have a power generator at a constant speed (in so far as the opposing torque is kept constant) for the regulated case [40]. For N lying between 0 and 1, the speed regulation is better than the one given by a flywheel. For N lying between 1 and 2 we also get a better regulation which can be considered as an overcompensation because the speed of the shaft for the regulated case is increasing. For values of N greater than $N = 2$ the speed of the shaft for the regulated case increases with a rate greater than the rate of decreasing speed of a flywheel.

We now consider the case where the device is supposed to compensate torque fluctuations due to variations of the load connected to an electric motor. We always assume that the mean value of the torque is provided by the electric motor. Fig. 10 illustrates a hydraulic control device and Fig. 11 the corresponding block diagram. (Equations [41] and [43] still give the result.)

It seems useful to investigate the kind of speed regulation which can be expected with a passive control device. The simplest one consists of springs that attract the balls toward the shaft. We choose a linear spring, the characteristic of which is

$$F_r = -kr \qquad [48]$$

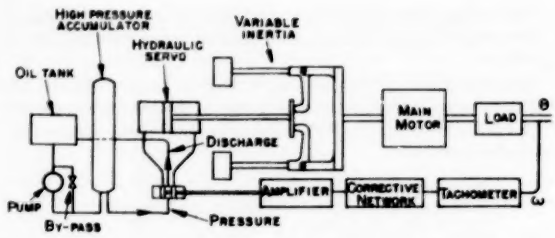


Fig. 10 Regulation of Speed by a Controlled Variable Inertia

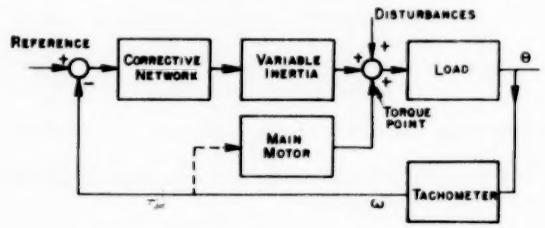


Fig. 11 Block Diagram of Fig. 10

The general equations of the system (see Equations [29] and [30]) are

$$(I_1 + mr^2)\,\frac{d\omega}{dt} + 2m\omega r\,\frac{dr}{dt} = C \qquad [49]$$

$$m\left[\frac{d^2r}{dt^2} - r\omega^2\right] = -kr \qquad [50]$$

We shall assume that

$$k = m\omega_0^2 \qquad [51]$$

ω_0 being the initial speed. Initially the balls are then in relative equilibrium and we assume that at $t = 0$ we have $r = R$. Equation [50] can be written as

$$\frac{d^2r}{dt^2} - r\,(\omega^2 - \omega_0^2) = 0 \qquad [52]$$

As before, we assume that $I_1 = 0$ (to simplify the calculations). The system [49] and [50] cannot be directly integrated, although a first integral of [49] can be written as follows

$$mr^2\omega = Ct + mR^2\omega_0 \qquad [53]$$

Letting

$$C = mC_1 \qquad [54]$$

we have to solve

$$\begin{cases} r^2\,\omega = C_1\,t + R^2\omega_0 & [55] \\ \dfrac{d^2r}{dt^2} - r(\omega^2 - \omega_0^2) = 0 & [56] \end{cases}$$

We make the transformation

$$r^2 = z \qquad [57]$$

and after elimination of ω we obtain the equation

$$\frac{1}{2}\,zz'' - \frac{1}{4}\,z'^2 + \omega_0^2 z^2 - (C_1\,t + R^2\omega_0)^2 = 0 \qquad [58]$$

The solution $r = r(t)$ of this equation has been obtained from an electronic-analog computing machine.[10] The solution which corresponds to the following numerical values:

$$\omega_0 = 7.5 \text{ rad/sec}$$
$$C = mC_1 = 10^7 \text{ ergs} = 1 \text{ joule}$$
$$m = 3 \text{ kg}$$
$$R = 0.3 \text{ m}$$

is represented by Fig. 12. From it we have deduced the curve $\omega(t)$

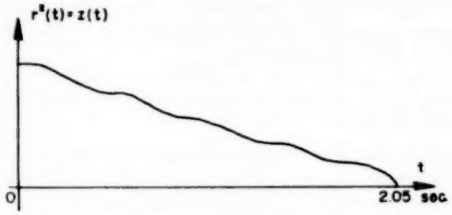


FIG. 12 SOLUTION OF EQUATION [58] FOR PARTICULAR NUMERICAL VALUES

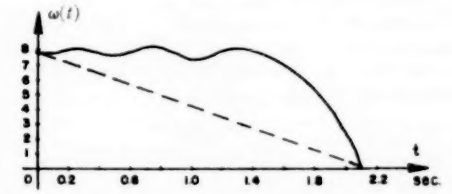


FIG. 13 ANGULAR SPEED OF SELF-ADJUSTED VARIABLE INERTIA CONNECTED TO CONSTANT LOAD

(Equation [48]) by step-by-step computation. The curve $\omega(t)$, plotted in Fig. 13, shows that in a first approximation the system leads to an acceptable speed regulation for 1.4 sec. We note that the system must be stabilized because it tends to oscillate. Stabilizing devices are being studied.

REGULATOR SUBJECTED TO STATISTICAL DISTURBANCES

We suppose that the opposing torque is no longer constant. Let $c(t)$ be the torque which is assumed to be a stationary random function. If C is the mean value of that torque we then consider $c(t)$ so that $\overline{c(t)} = 0$ and

$$C(t) = c(t) + C \quad [59]$$

We suppose that the autocorrelation function $\gamma(\tau)$ of $c(t)$ is known, namely

$$\gamma(\tau) = \overline{c(t)\,c(t+\tau)} \quad [60]$$

or its spectrum

$$\Gamma(s) = \int_{-\infty}^{+\infty} \gamma(\tau) e^{-js\tau}\, d\tau \quad [61]$$

The problem consists of determining the characteristics of the regulator in such a way that the angular speed of the shaft $\omega(t)$ is as constant as possible. The mean-square error is

$$\overline{[\omega(t) - \omega_0]^2} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{+T} \omega^2(t)dt - \omega_0^2 = \bar{\omega}(0) - \omega_0^2 \quad [62]$$

[10] "Laboratoire de Calcul Electronique" of the "Service Technique Aéronautique," Paris. We thank Mr. Decker who authorized us to put the problem on the computer.

where ω_0 is the mean value of $\omega(t)$, and $\bar{\omega}(\tau)$ is the autocorrelation function of $\omega(t)$, namely

$$\bar{\omega}(\tau) = \overline{\omega(t)\omega(t+\tau)} \quad [63]$$

We can envisage the solution of all the cases already considered in the previous section. However, computations are very lengthy and because of space limitation we can treat only one case; we shall state the problem and indicate how to solve it (some computations are being made at present on an electronic computer).

In order to judge whether or not the variable-inertia regulator is worth while we must compare it to the constant-inertia wheel, whose inertia is equal to the maximum inertia of the regulator. Let I be the inertia of such a wheel. We have

$$I \frac{d\omega}{dt} = c(t) \quad [64]$$

From Equation [64] (and the one we can get when t becomes $t + \tau$, where τ is a parameter) we obtain

$$I^2 \left(\frac{d\omega}{dt}\right)_t \left(\frac{d\omega}{dt}\right)_{t+\tau} = c(t)\,c(t+\tau) \quad [65]$$

Taking the mean value of the two members, we get

$$I^2 \overline{\left(\frac{d\omega}{dt}\right)_t \left(\frac{d\omega}{dt}\right)_{t+\tau}} = \gamma(\tau) \quad [66]$$

The left number is related to the autocorrelation function[11] $\bar{\omega}(\tau)$ of $\omega(t)$. This is $-I^2\bar{\omega}''(\tau)$.

We then have

$$-I^2\bar{\omega}''(\tau) = \gamma(\tau) \quad [67]$$

We obtain the expression for the power-density spectrum $\Omega(s)$ of $\omega(t)$ noting that derivative in the time domain corresponds to multiplying by s in the complex domain

$$I^2 s^2\, \Omega(s) = \Gamma(s) \quad [68]$$

from which

$$\Omega(s) = \frac{1}{I^2} \frac{\Gamma(s)}{s^2} \quad [69]$$

We then know the output spectrum (speed spectrum) from the input spectrum (torque-fluctuation spectrum). If we also assume that the probability distribution function of $c(t)$ is known, i.e.

$$F_1(\underline{c}) = \text{prob } c(t) \leqslant \underline{c} \quad [70]$$

(where prob means "probability for having"), we can compute the probability distribution function of the speed $\omega(t)$ in the case of a constant-inertia wheel. From Equation [64] we have

$$\omega - \omega_1 = \frac{1}{I} \int_0^t c(u)du \quad [71]$$

The repartition function $F_2(\underline{\omega})$ of $\omega(t)$ is [see reference (9)]

$$F_2(\underline{\omega}) = \text{prob } \omega(t) \leqslant \underline{\omega} \quad [72]$$

[which is related to $F_1(\underline{c})$].

[11] If $\varphi(\tau) = \overline{f(t)f(t+\tau)}$
we have

$$\varphi'(\tau) = \frac{d\varphi(\tau)}{d\tau} = \overline{f(t)f'(t+\tau)}$$

$$\varphi''(\tau) = \frac{d^2\varphi(\tau)}{d\tau} = -\overline{f'(t)f'(t+\tau)}$$

See reference (8).

We now come back to a controlled-inertia regulator, the controlling energy coming from an external source. We start with the equation

$$(I_1 + mr^2)\frac{d\omega}{dt} + 2mr\omega\frac{dr}{dt} = c(t) \quad [73]$$

We set $I_1 = 0$. We then have the first integral

$$mr^2\omega = \int_0^t c(u)du + mR^2\omega_0 \quad [74]$$

The most satisfactory law to be imposed on $r(t)$ would consist in relating $r(t)$ to the random torque $c(t)$. The difficulty does not arise from the fact that $c(t)$ is a random function but rather from the fact that it is difficult to control a quantity from a torque. A control device for which $r(t)$ is chosen to satisfy

$$mr\frac{dr}{dt} = kc(t) \quad [75]$$

where k would be a parameter to be determined, should be satisfactory because Equation [49], with its two members multiplied by ω, shows that the work furnished by the system and the work absorbed are equal.

We then try, as a first approximation to the regulation of speed, to control the position of the balls from a quantity which can be more easily detected, namely, the actual speed of the shaft. We assume that we adjust $r(t)$ so that

$$\frac{dr}{dt} = \alpha\frac{d\omega}{dt} \quad [76]$$

is satisfied where α is a parameter to be determined. Then

$$R - r = \alpha(\omega_0 - \omega) \quad [77]$$

Equations [74] and [77], after eliminating r, give

$$m\omega[R - \alpha(\omega_0 - \omega)]^2 = \int_0^t c(u)du + mR^2\omega_0 \quad [78]$$

which is a third-order equation

$$m\alpha^2\omega^3 + 2m\alpha(R - \alpha\omega_0)\omega^2 + m(R - \alpha\omega_0)^2\omega - \int_0^t c(u)du - mR^2\omega_0 = 0 \quad [79]$$

The product of the roots is positive, which means that if there is only one real root this one is necessarily positive. The term $mR^2\omega_0$ is, as a matter of fact, always greater than $\int c(u)du$ because the left number of Equation [74] is positive.

It is now easier to solve this equation with numerical values[12] than symbols. We compute the root $\omega(t)$ from Equation [78] or better, the square of that root $\omega^2(t)$. It involves the integral $\int_0^t c(u)\,du$; it is then of the following form

$$\omega^2(t) = f(\alpha, c(t)) \quad [80]$$

We take the mean value of both members of this equation. We then get the mean-quadratic value of the speed $\omega(t)$. We now come back to Equation [79] and average its members. The system must be such that

$$\overline{\omega(t)} = \omega_0 \quad [81]$$

If not, it could not be considered as stable. Consequently, after averaging, Equation [79] can be written as

$$m\alpha^2\overline{\omega^3} + 2\alpha m(R - \alpha\omega_0)\overline{\omega^2} + m\alpha^2\omega_0^3 - 2mR\alpha\omega_0 = 0 \quad [82]$$

from which we deduce $\overline{\omega^3(t)}$.

We thus obtain the first three coefficients $\overline{\omega(t)}$, $\overline{\omega^2(t)}$, and $\overline{\omega^3(t)}$ of the expansion in powers of the characteristic function $\phi_3(u)$ of the repartition function $F_3(\omega)$ for $\omega(t)$ and the variable-inertia wheel. The comparison of these terms with those corresponding to $F_2(\omega)$ (for the constant-inertia wheel) allows us to judge the efficiency of the controlled variable inertia and to adjust its parameter by choosing a relation such as Equation [76].

For that particular problem we note that the comparison between the mean-square errors is sufficient to determine the optimum value of α in Equation [76].

Until now we have not considered the technical feasibility of the problem. It seems to us that two main features should be studied carefully:

1 The balancing of the device because it is not a rigid system. We are investigating the possibility of using a passive self-centering system. Such a system is theoretically possible.

2 In the case of the self-controlled variable inertia the energy initially stored in the spring is equal to the total kinetic energy of the wheel. We then need one or several springs with very high rigidity and also a large stroke.

On the other hand, it seems possible to replace the balls by a fluid (mercury) and a pump. The over-all efficiency would probably be less than the one with the mechanical device but the realization would certainly be easier.

Conclusion

The purpose of this paper has been to call the attention of the reader (*a*) to our recent work on the introduction of the probability calculus into the servomechanisms field; (*b*) we have tried to illustrate these theories by an example. We have not introduced a new device, but we feel that we have brought it up to date by a new application, namely, the variable-inertia regulator.

It should not be concluded that the application of such a device is in the distant future. Let us recall that some years ago an auxiliary take-off device for airplanes was constructed by Westinghouse. It consisted in the use of the kinetic energy of a flywheel which then was connected to an alternator. Local cars or buses powered by energy stored in a flywheel have been—and still are—studied and realized. The generation of electric power in guided missiles appears to offer the widest field of application.

As to the use of statistical computation, let us take the following example: A turbine is connected to an alternator and an electrical network. We can imagine a variable-inertia wheel locked on the common shaft, the purpose of which is to control the frequency of the output current in such a way that it stays a constant in spite of the variation of load. In this example the statistical properties of the load disturbances are known a priori. If the network is large enough, the probability distribution of the intensity follows a Gaussian law (addition of a great number of independent random variables, reference 3) and the instants at which changes in values appear follow a Poisson distribution (see Equation [19]) because the probability that a change will occur in an interval of time dt is kdt where k is the average number of changes per second.

Such examples are rare, but as soon as the stationary property of the phenomenon is established it is a simple matter to record the phenomenon and to study its statistical properties at leisure. The need of an automatic computing machine[13] is unavoidable.

[12] Calculations are in process.

[13] Use of punched-card machines is the cheapest way for general studies. But for extensive use of these theories special machines have to be designed.

ACKNOWLEDGMENT

The author is indebted to Dr. Y. W. Lee for the initiation into statistical theories, Profs. J. Pérès and G. Darmois for the supervision of the work, and to Dr. R. Oldenburger for his revisions of the manuscript.

BIBLIOGRAPHY

1 "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948.
2 M.I.T. Course: "Statistical Communication Theory," by Y. W. Lee, class note (a book on this to be published).
3 "Généralité sur les Probabilités. Éléments aléatoires," by M. Frechet, Gauthier, Villars, Paris, France, 1950.
4 "The Intrapolation, Extrapolation and Smoothing of Stationary Time Series," by N. Wiener, John Wiley & Sons, Inc., New York, N. Y., 1949.
5 "Les Fonctions Aléatoires," by S. Bernstein, E. Slusky, and A. Steinhauss, Actualités Scientifiques et Industrielles No. 738, Hermann, et Cie, Paris, France.
6 "Calcul Statistique des Systèmes Asservis," by M. Pélegrin, Thesis, November, 1952.
7 "Brevet d'Invention État Francais (Invention, M. Pélegrin)," French Patent No. 612.377, June 15, 1951.
8 "Some Analytical and Practical Aspects of Weiner's Theory of Prediction," by Robert Cohen, Massachusetts Institute of Technology, Cambridge, Mass., Technical Report, No. 69, p. 31.
9 "Statistic Processes," by J. L. Doob, John Wiley and Sons, Inc., New York, N. Y., 1953, chapter 1.

Discussion

R. OLDENBURGER.[14] This paper is particularly important in that the author is concerned with the direct industrial application of Wiener's statistical approach. To make the results of the paper available and understandable to the reader wno is not an expert on Wiener's theory, parts of the paper need considerable clarification. Because his thesis is virtually inaccessible to the members of this Society it will be very helpful if the author in his closure will explain the introduction of the function $P(s)$, why $Q(s)$ must have its poles in the right half plane only, why $H(s)$ must tend to zero as $1/r^2$, and, in fact, give that part of the argument of his thesis that applies to his paper.

Equation [32] of the paper comes from the law that the torque is equal to $d(I\omega)/dt$ for the total moment of inertia I. In his study the inertia I may be a variable. As a consequence, we must use

$$\frac{d(I\omega)}{dt} = I\frac{d\omega}{dt} + \omega\frac{dI}{dt},$$

to obtain Equation [32].

In view of Equation [33] the rate at which work is done in moving the balls is given by

$$mr\omega^2\frac{dr}{dt} \quad \ldots\ldots [83]$$

The Term [83] is the power associated with the movement. It is now assumed that the energy used to move the balls is drawn from the rotational energy of the system. An efficiency η is assumed for this energy conversion, whence the power drawn from the rotating member is

$$\mu mr\omega^2\frac{dr}{dt} \quad \ldots\ldots [84]$$

where $\mu = 1/\eta$. It is to be recalled that the product of torque and angular speed is power. Treating the power given in Expression [84] as one that comes from a torque associated with the rotating member, one obtains the equivalent torque by dividing the power [84] by ω. This accounts for the last term in Equation [35] of the paper.

Reference (8) of the paper is not readily accessible to most readers of the paper. The author in his closure can furnish the reader valuable aid by citing a reference to a book or journal of general international circulation or by giving the necessary mathematical background for the section of his paper entitled, "Regulator Subjected to Statistical Disturbances." The argument concerning "work" and Equation [75] is not clear to the writer. Does the author assume in Equation [73] and what follows that ω is positive? Even though ω cannot be negative physically, the writer does not see why ω cannot be negative in the solution of the mathematical formulation of the problem. Clarification of this point in the author's closure is needed and it will be most helpful to have the correct, rigorous, though delicate, mathematical proof that the mean value of

$$\int_0^t c(u)du$$

vanishes (or a readily accessible reference where the subject is treated in thorough detail). The meaning of Equation [82] would be much more easily understood if the author's closure would cover the expansion of $\phi(u)$ in terms of $\{M_n\}$ in detail so as to clarify the sense of the material following Equation [82]. Such an explanation was in the original manuscript. The author removed it at the suggestion of the reviewers. This was an unfortunate abbreviation of the paper which the author can rectify in his closure.

The Society is fortunate that the author has made available the results of the enormous amount of work he had to undertake in the computation of the statistical functions of Wiener's theory. This tremendous amount of work is inherent in applying Wiener's theory and is a major obstacle to its use in industry.

EUGENE W. PIKE.[15] The author communicates, with warming enthusiasm, the results of his early (and very promising) work on servomechanism theory. His paper contains some minor errors, however, which should be noted for the benefit of those who may make their first contact with the field in this paper.

The response of servomechanisms to statistically defined inputs and disturbances is a well-developed subject with a tremendous literature. In this writer's opinion, the extreme generality and power of Wiener's own very abstract treatment, and the book of James, Nichols, and Phillips in the Radiation Laboratory series (the exact reference is not at hand) should have been mentioned in the introduction.

In the general Wiener theory, the optimum signal-restoring filter is defined not only by the autocorrelation functions of the signal and the noise or disturbance, but also by the time lag allowed in the filtering process. The author's treatment apparently deals with zero time lag (his Equation [17]) but this is not clear from his text.

That spectra derived from autocorrelation functions are approximable by rational functions is not generally true in practice. To give an extreme example, suppose the disturbance function θ_p in his Fig. 3 is a randomly intermittent 60-cycle hum. This has a well-defined autocorrelation function, its spectrum is a narrow spike at 60 cycles, and the rational function required to provide a safe approximation for design purposes might require hundreds of factors. This whole area of approximating power spectra by one or another class of functions can be treacherous.

The whole of the theory of the variable-inertia rotor is greatly simplified if one approaches it from the expression for the conservation of angular momentum

[14] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

[15] Head, Applied Physics Section, Raytheon Manufacturing Company, Waltham, Mass.

$$\frac{d}{dt}(m r^2\omega) = C(t)$$

R. C. Booton, Jr.[16] This paper is an interesting description of some of the applications of statistical techniques that have been undertaken in France. The author probably should have emphasized that this was his aim rather than making any significant contribution to the application of statistical techniques. It is unfortunate that no numerical results were presented for the statistical analysis of the relays and of the variable-inertia regulator since the casual reader will wonder whether it is possible to obtain such results.

Author's Closure

In reply to Dr. Oldenburger's comments I want to go back to general remarks made at the beginning of my paper. The calculations referred to in my paper and developed in my thesis were mentioned only to support the conclusions. I doubt that it is possible to derive Equation [18] directly without several pages of computation, but as soon as the equation is obtained, its application to a particular problem does not need any longer or more complicated computations. The solution of a problem can be derived directly from Equation [18] and it is not necessary to go back to its proof each time a problem occurs. In other words, this is the starting point of the derivation of the solution.

Now, I will try to answer Dr. Oldenburger's comments. A Laplace or Fourier transformation transforms a function $f(t)$ into a function $F(\omega)$. If $f(t)$ is submitted to certain conditions it may result in certain corresponding conditions on $F(\omega)$. This is the case for a Fourier transformation acting upon a function $f(t)$ which has a nonzero value only for $t > 0$ or $t < 0$. The "transformation of this condition" is that the function $F(\omega)$ has "no poles in the left or right half plane." The computation leads directly to the condition that $P(s) \cdot [H(s) - 1]$ must have no poles in the left half plane [$Q(s)$ is equal *by definition* to the foregoing expression and is used for abbreviation].

As I have said previously, this is the "transformation" of a condition which applies in the time domain to an expression which has a nonzero value only for $t > 0$. I could not develop the details of the computation. It is a straightforward application of the calculus of variations (see 6). I take this opportunity to emphasize that reference 6 is commercially available.[17]

As to the expression for $P(s)$, I cannot establish it here as it would take too long a time. Let me say only that it derives directly from the Fourier transformation of the time domain equations of the system.

Why must $|H(s)|$ tend to zero as $1/r^2$? It is to be remembered that we restricted the solution to *rational solutions;* this means that fractional or irrational powers of s do not appear in $H(s)$. Moreover, $H(s)$ is still a transfer function of a physical system

$$H(s) = \frac{A \cdot B}{1 + A \cdot B}$$

Its Laplace transform, $h(t)$, which is the response of the system to an impulse placed at $t = 0$, must be so that $h(t)$ tends to zero when t tends to infinity. The "final value theorem" gives

$$\lim_{t \to \infty} h(t) = \lim_{t \to 0} sH(s) \quad [85]$$

If $sH(s)$ must tend to zero, this means that $H(s)$ is of the order of $1/r^2$ ($r = |s|$) or more ($1/r^n$ with $n \geq 2$).

Concerning the point of Equation [32] mentioned by Dr. Oldenburger, some additional details are given in the answer to Dr. Pike's comments.

We come to the sign of ω. From the physical considerations we can consider only the case where $\omega > 0$. The answer from a purely mathematical viewpoint makes no sense and, in any case, r needs careful definition. The axis along which r is computed must be orientated. It is obvious that if ω becomes zero when $r = 0$ the device could no longer furnish any energy. Moreover, we can restrict our solution to $r > 0$ because the case $r < 0$ corresponds to a case $r > 0$ after half a turn. We can then always find an initial value R so that the first member of Equation [74] is positive. In that case ω will always be positive.

The mathematical proof of the mean value of

$$\int_0^t c(u)du$$

is very delicate. On the other hand, the engineer's proof is quite evident and does not seem to require much emphasis. An intermediate explanation can be derived as follows.[18] Let us call $\theta(a)$ the mean value with regard to t of

$$\int_a^t c(u)du \quad [86]$$

This is

$$\theta(a) = \overline{\int_a^t c(u)du}^{\,t} \quad [87]$$

The subscript t is put on the bar to specify the variable with regard to which the mean value is taken. We have

$$\int_a^t c(u)du = \int_0^t c(u)du - \int_0^a c(u)du \quad [88]$$

We derive

$$\theta(a) = \theta(0) - \overline{\int_0^a c(u)du}^{\,t} \quad [89]$$

The last integral no longer depends on t. We then can write the foregoing equation

$$\theta(a) = \theta(0) - \int_0^a c(u)du \quad [90]$$

If we now take the mean value of $\theta(a)$ with regard to a we get

$$\overline{\theta(a)} = \theta(0) - \theta(0) \quad [91]$$

The mean value of the second term on the right of Equation [90] is, by definition, $\theta(0)$.

Concerning the last question in Dr. Oldenburger's comments, I wish to recall that the characteristic function $\phi(u)$ of a repartition function is

$$\phi(u) = \int_{-\infty}^{\infty} e^{iux} dF(x)$$

From the series expansion of

$$e^{iux} = 1 + \frac{iux}{1!} + \frac{i^2u^2x^2}{2!} + \ldots + \frac{i^nu^nx^n}{n!} + \ldots$$

we get, by substituting the foregoing expression in the definition of $\phi(u)$ and integrating each term

$$\phi(u) = 1 + i\frac{\mathfrak{M}_1 u}{1!} + i^2\frac{\mathfrak{M}_2 u^2}{2!} + \ldots + i^n\frac{\mathfrak{M}_n u^n}{n!}$$

[16] Research Engineer, Dynamic Analysis and Control Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

[17] Apply to Ministére de l'Air, Service de Documentation et d'Informations Techniques, 4, Avenue de la Porte d'Issy, Paris, 15e, France, and ask for P.S.T. No. 285, price 1,332 francs, postage included.

[18] Discussed with Professor Kuntzman, Grenoble, France.

Here the $\mathfrak{M}_j$ is the moment of the j-th order of the random function $\theta(t)$

$$\mathfrak{M}_j = \int_{-\infty}^{\infty} \theta^j(x)p(x)dx$$

The expansion in series of $\phi(u)$ just written is formal and does not necessarily converge.

A good available book is the one recently published in France, "Théorie des fonctions Aléatoires," Blanc-Lapierre et Fortet, Masson Editeur, Paris, France.

In reply to Mr. Booton's comments, I wish to say that my paper is not a "description of some of the applications of statistical techniques that have been undertaken in France" but only one of the work that I have undertaken. Obviously, there are many other workers in this field in France as in any other country.

Concerning the last point, let me say that I have built a variable-inertia regulator (see Fig. 14). Experiments, which have been preceded by numerical computation, are in process at present.

In reply to Dr. Pike's comments I emphasize that the question of "time lag" is very important, and I thank Dr. Pike for giving me the opportunity to point out the assumptions. In the general case (i.e., with time lead or time lag) the Wiener theory leads to very long computations (they are developed in reference 6). The present calculations are very simple only because I have assumed there is no lag or lead time. Such an assumption is valid because it is in the servo field. Let us compare this field to the communications field.

A time lag in the transmission process of a message does not present serious objections and sometimes permits improvement in the message (message plus noise). This is a permanent process. On the other hand, consider a servo, the purpose of which is to drive a gun toward a (corrected future) target. The gun must be pointed at the target at *any time* in order that, if other conditions are satisfied, the shell can be fired. This particular example covers, in fact, a large number of servos.

As to the possibility of approximating an autocorrelation function by a rational function, I wish to say that it is less difficult than it seems at first.

If there exists a "60-cycle hum" in the signal (noise) with a very sharp spectrum, this means that the frequency of the hum is well stabilized and does not fluctuate around its mean value. In that case a conventional narrow-band-pass filter, with a very high resonant factor, will eliminate the noise quite well. If the frequency of the hum is not well stabilized, this means that the hum is rather a noise and as a consequence its spectrum is no longer so sharp. The case with which one can approximate it by a set of orthogonal functions is improved. Moreover, however sharp the function, there exist expansions in series of functions which permit good approximation.

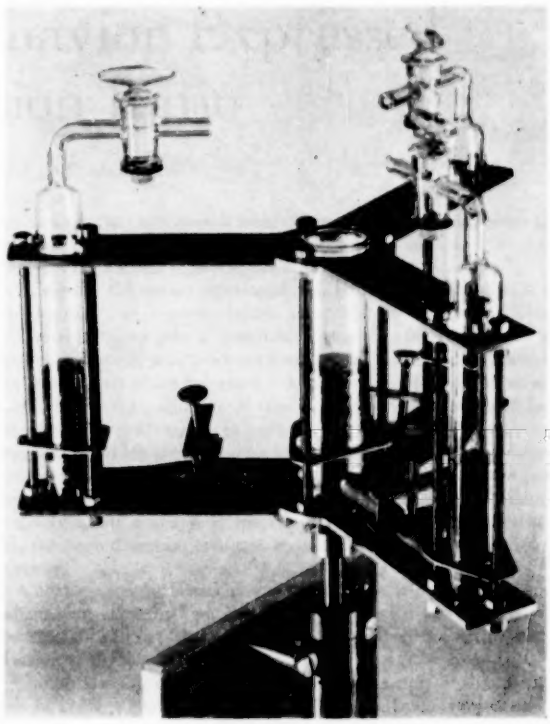


FIG. 14 VARIABLE-INERTIA REGULATOR

The use of automatic computers is necessary; the computation of coefficients of an expansion in series (of orthogonal functions, for example, as suggested in the text) can be made with an autocorrelator or any analog computer (low-price computer) with an automatic curve-follower.

Concerning Dr. Pike's last comment, and Dr. Oldenburger's comment, I would point out that Equation [29] is exactly the same as the one mentioned by Dr. Pike at the end of his comments. The use of the conservation of angular momentum, which is Equation [29], does not tell anything concerning the *radial forces acting upon the balls*. See, for example, Equation [50]. Therefore, because I need the expression of the radial forces I prefer to present the solution in a more logical form in giving the expression of the two components of the acceleration directly. The conservation of the angular momentum is only the "first integral" of Equation [29] and does not completely solve the problem. See text above Equation [53].

Control-System Behavior Expressed as a Deviation Ratio

By J. M. L. JANSSEN,[1] DELFT, HOLLAND

In order to bridge the gap between process-control practice and modern control theory, an attempt is made to reduce this theory to its bare essentials. A simple picture is drawn of what control does and how it does it, use being made of the familiar concepts of resonance and damping. Thus a theoretical background is provided for common-sense judgments on control quality and controller settings. The picture given is applicable to both feedback and feedforward control.

Introduction

IN many branches of engineering there is a gap between theory and practice. On the one hand, for instance, theorists feel that practical men do not always take advantage of their results. On the other hand, practical men consider theory to be often unrealistic, and thus are not always receptive of its results; moreover, in many cases a practical answer has been found long before theory has come into play.

In process control too this situation prevails. Good practical results have been achieved without much theoretical support. When in the related field of servomechanisms elegant theories based on frequency-response techniques came into use, quite naturally attempts were made to use the same methods in the field of process control, and there was the gap and the lack of response.

This paper is intended to contribute toward bridging this gap and it is written in the belief that some of the main causes of the lack of response are not too difficult to identify.

One cause is inherent in the theoretical approach. Theory is bound to work with a scheme that is an abstraction and a simplification of reality. A theorist often forgets that reality is much more complex and, consequently, he often presents his results as "the" solution of the problem, when he has solved only that part of it that lends itself to quantitative studies. He also sometimes feels inclined to think that practice should be made conformable to the simplifications he made for his convenience which, however, may be very inconvenient from the practical point of view.

Another cause is that modern theory of control originates in a field that is foreign to people, such as chemical engineers, concerned with the operation of processes. It does not answer to their frame of mind. Frequency-response methods, which come quite naturally, for instance, to communication engineers, seem rather artificial to their colleagues in the chemical-engineering field.

A third cause consists in theory often not being modest enough. In process control a novel theory is applied to an already well-established field. Then, the first point on the program should be, not to propose revolutionary reforms, but to grasp why things are done as they are, even if from a theoretical angle they seem to be done in the wrong way. Only if theory is successful in this respect may success in new applications be expected.

In view of the causes mentioned, this paper is concerned with a "philosophy" of control, rather than with a "theory." This will enable us to give a qualitative over-all picture of control system behavior, which seems more useful than a quantitative picture of part of its behavior. A distinction is made between what control does and how it does it, the first of which is of interest to chemical engineers concerned with process operations, and both of which are of interest to people concerned with process control. Accordingly, the first two sections, dealing with what control does, are in simple terms, avoiding all mathematics, making use only of the concepts of resonance and damping. In particular, the second section analyzes some aspects of common control practice.

A next section deals with how control works, using only graphical means. The final section is concerned with the theoretical justification of part of the material presented in the earlier sections.

The philosophy presented here has its proper field of application in the automatic control of plants in continuous operation. For automatic start-up, for instance, other methods are more suitable.

What Automatic Control Does

One of the problems with which we are faced in technical work is how to express in exact figures the quality of some product. A similar problem arises if one tries to define a more abstract concept like control quality. Various authors have contributed to condensing into concise formulations the vague ideas that are used more or less unconsciously. Ziegler and Nichols (1),[2] for instance, in their classical paper on "Optimum Settings for Automatic Controllers," use speed, stability, and deviation reduction as criteria for control quality. Rutherford (2) uses the same criteria, but in a quantitative way. His method of measuring deviation reduction will be considered in detail, and will be made a starting point for further discussions. He introduces a "deviation reduction factor," which is the ratio between the potential deviation and the peak value of the measured deviation, Fig. 1.

By working with potential deviations, all possible disturbances, whatever their character, are reduced to directly comparable figures. This does not mean that the value of the deviation reduction factor is independent of the character of the disturbances. For a step change in set point, for instance, the deviation reduction factor equals unity, since at the moment of applying the disturbance the deviation equals the step change. However, if a step disturbance is injected somewhere in the plant, then the deviation reduction factor will be greater, owing to the smoothing of the disturbance on its way through the plant to the controller; this is shown by the dash line in Fig. 1.

Obviously, the solid line in Fig. 1 depends not only on the ultimate value of the dash line (potential deviation), but also on its entire course.

[1] Research Engineer, Royal-Dutch/Shell-Laboratory.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 19, 1953. Paper No. 53—A-21.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

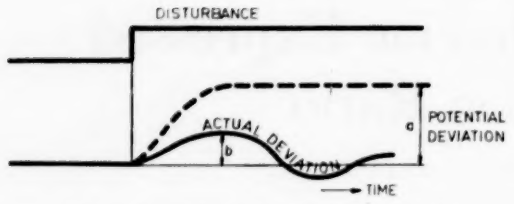


FIG. 1 DEVIATION-REDUCTION FACTOR = a/b

This course is determined by plant properties and the character of the disturbance. In other words: the dash line shows what the plant does to a specific disturbance. It is independent of what control does. It is not suggested here that control is independent of plant properties, but only that the plant also has an effect of its own.

Both elements, "what the plant does" and "what control does," are of interest. The deviation reduction factor combines them, which has the advantage of giving direct information about the effect of a specific disturbance, including the influence of plant dynamics on this effect.

Keeping the elements apart and considering first only what control does, has the advantage that no specific information is needed about disturbances. Only their effects are of interest: the actual deviations, which usually are recorded on the chart of the controller. Once a plant is in operation this information is available. In the design stage things are different. Then what the plant does has to be considered in addition. The same is true if steps are to be taken to deal with the causes of the actual deviations.

From what follows, it will appear that this two-step approach—what does the plant do—what does control do—enables a simple picture of control-system behavior to be drawn.

In this paper the second step will receive most of the attention. The first step is harder to deal with in a general way, and many other aspects besides those of control have to be considered.

To determine what control does, the actual deviations when the plant is under control can be compared with the actual deviations when the controller is put out of operation. However, the result of this comparison cannot be expressed in a single figure, such as, for instance, the ratio of some average value of the deviations with and without control. If this were attempted, a combination of properties would be measured, the ratio being dependent on both the behavior of the control system and the time pattern of the actual deviations without control. Moreover, it would be a disadvantage that by taking averages all information concerning time patterns would be forfeited. In some cases this does not matter, but in other cases it is useful to know whether the remaining deviations, although small, are rapid or slow.

It is possible to split the combination of "time pattern" and "what control does" by specifying the time pattern of the deviations, and it is an obvious course to use sine waves for this purpose in order to tie in with modern theories of automatic-control systems. Then, not only how control works, but also what it does, can be dealt with by frequency-response methods. Thus for a single sinusoidal deviation a "deviation ratio" can be defined as follows:

$$\text{Deviation ratio} = \frac{\text{Deviation with control}}{\text{Deviation without control}}$$

the only difference between the two situations "with" and "without" being that in the latter case the controller is disconnected.

For deviation one may read either "actual" or "measured" deviation, if only the same adjective is used in numerator and denominator. Replacing actual by measured amounts to changing numerator and denominator in the same proportion, and so accounting for the attenuation in the measuring means. Thus unlike the deviation reduction factor the deviation ratio has no "masking effect" (3).

Of late years this deviation ratio has been the central concept in our philosophy of automatic process control and of our fairly extensive educational activities in this field. In the meantime, we have learned that Ahrendt and Taplin (4) have been using the same concept, but it has not been worked out to the same extent in their book as it has in our studies on process control. They use the term "error ratio," which, in recognition of their prior work, we would adopt, but that the term "deviation" is in better agreement with recent proposals for process-control terminology (5).

Deviation Ratio in Feedback Control. As Ahrendt and Taplin (6) show, the deviation ratio can be derived in a simple way from known properties of the elements constituting the control system. Yet, in our opinion, even this possibility of simple quantitative treatment of control problems, much as it means, is only the minor gain. We have found the greatest advance to lie in the simple qualitative over-all picture of control-system behavior that develops from the concept of deviation ratio. This is particularly useful because this picture clearly reveals one essential aspect of control-system behavior that remains somewhat obscured if methods based on transient response are used, namely, the phenomenon of resonance.

It is our experience that control engineers seldom realize that automatic feedback control must cause amplification of the effects of certain types of disturbances instead of the reduction expected. They find it especially surprising in cases where transients are overdamped. This possibility of amplification, however, is at once apparent from Fig. 2 which shows a deviation-ratio curve measured for a flow-control circuit.

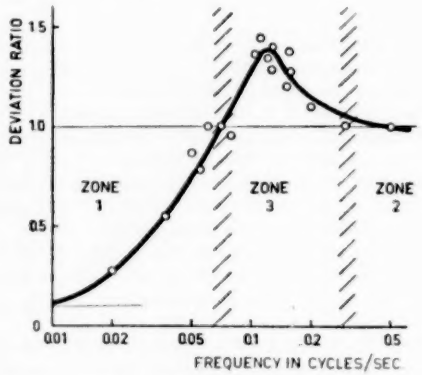


FIG. 2 DEVIATION-RATIO CURVE OF AN EXPERIMENTAL FLOW-CONTROL CIRCUIT

Generalizing this picture, one may state that the entire frequency range can be subdivided into three zones:

1 A zone of low frequencies where there is true deviation reduction.

2 A zone of high frequencies where control is practically inoperative.

3 A zone of medium frequencies where resonance occurs and the effect of disturbances is amplified.

This "three-zone" picture seems to us the simplest one that shows all the essential elements of control-system behavior. Moreover, it is simple to explain. In zone 1 the plant responds to the regulating valve fast enough to enable proper control. In the high-frequency zone the plant is too slow to give any

appreciable response. For medium frequencies there is a response, but it is lagging, and near the resonance frequency the lag is such that the correction applied has the wrong direction, thus amplifying the original deviation.

To give some idea of the behavior of the deviation ratio in the resonance zone: The amplification in the resonance peak is more than twofold for controller settings according to Ziegler and Nichols' rules. This is the price that has to be paid for obtaining deviation reduction at low frequencies. It may be asked whether this resonance could not be avoided by designing control systems more cleverly. It has been shown by Westcott (7) that it is basically impossible to obtain any deviation reduction at some frequencies without paying for it by deviation amplification at other frequencies. He has found a simple relationship that governs the balance of profit and loss. In the next section Westcott's theorem will be dealt with in some detail.

Within the limitations set by Westcott's theorem, the deviation-ratio curve can be varied by varying the controller settings. In particular, the peak height is sensitive to such variations. The resonance frequency can be shifted, but it is less sensitive, process time lags being a major determining factor. In some instances it is possible to change the process time lag in the control loop by changing the location of the measuring and (or) manipulating means.

Fig. 3 gives a characteristic example of the way in which controller adjustments influence the shape of the deviation-ratio curve.

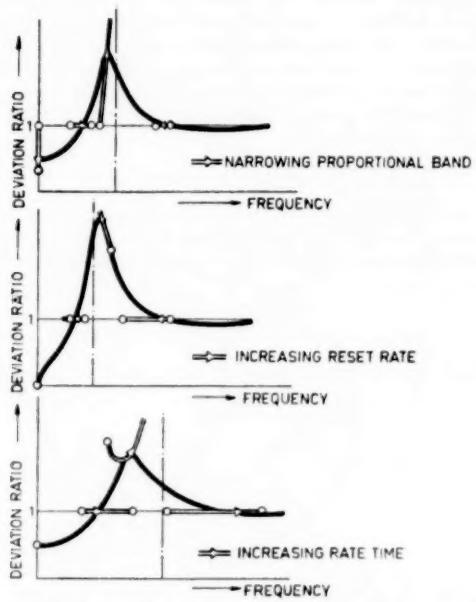


FIG. 3 EFFECTS OF CONTROLLER SETTINGS ON DEVIATION-RATIO CURVE

Deviation Ratio in Feedforward Control. So far, only automatic feedback control has been considered. The deviation ratio is also a useful tool for dealing with feedforward control, Fig. 4, for instance ratio control. The dynamic aspects of this type of control often are overlooked, perhaps because there is no possibility of hunting.

Farquhar (8) has given a thorough discussion on various types of feedforward control in terms of transient-response curves, and has stressed the importance of matching time lags. If these problems are dealt with in terms of frequency response, then there appears to be a possibility of resonance, just as in feedback systems. Again the explanation is simple. If in Fig. 4 the time lags in the paths P and Q are different, then at some frequencies the corrections will be applied in the wrong direction.

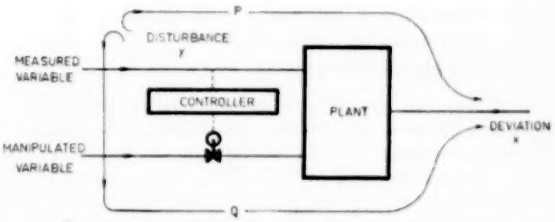


FIG. 4 FUNCTIONAL DIAGRAM OF FEEDFORWARD CONTROL

However, the deviation-ratio curve does not necessarily satisfy Westcott's theorem. It is, for instance, quite possible that a high-frequency variation in the measured variable is propagated with much less attenuation along path Q than along path P. In that case the deviation with control would be larger than the one without and, with increasing frequency, the deviation ratio would increase inordinately.

Often, however, the attenuation for high frequencies along path Q is the higher one, owing to the additional time lags in measuring means, controller, and valve. Then control is inoperative for high frequencies, the deviation ratio approaches unity with increasing frequency, and Westcott's theorem holds.

Summarizing, it may be stated that in many instances the three-zone picture is as adequate in feedforward control as in feedback control. In some cases, however, it fails.

In the case of failure considered, zone 2 and the resonance zone have amalgamated. The correspondingly high deviation amplification at high frequencies contributes to the debit side of the "balance of profit and loss," without there being a compensating contribution to the credit side.

REQUIREMENTS FOR DEVIATION REDUCTION

The deviation ratio expresses the capability of a control system to reduce the effect of a sinusoidal disturbance with a given frequency. In practice, disturbances are random, but they can be thought of as composed of sine waves. So in principle, for any sinusoidal component of the "deviation without control" the reduction owing to control can be determined. Then the question arises, what is the technical and economic benefit of such a reduction? If, for instance, in some applications rapid fluctuations do not cause any harm, then reduction of high frequencies is of little value, just as in cases where rapid fluctuations are absent from the deviation without control.

But that is not all. To complete the picture the valve movements should be taken into consideration. A control system cannot correct the effect of primary disturbances without introducing secondary disturbances. Take, for instance, furnace temperature control in an oil refinery. In order to correct temperature deviations in the heated oil stream the valve must manipulate the fuel stream. However, the corresponding fluctuations in heat supply are often undesirable from many other points of view, for instance from the point of view of furnace on-stream time, which is a nonlinear function of heat supply. In this example secondary disturbances are introduced into the furnace itself, but in other examples, owing to the interlinkages in complicated installations, they may be introduced into other parts of the plant as well.

So besides the deviation ratio the following factors also are of interest:

(*a*) Composition of deviations without control.

(*b*) Harm caused by deviations.

(*c*) Valve manipulations caused by control.

If all these factors could be taken into account quantitatively, then an optimum setting for the controller could be calculated. However, collecting the data and carrying out the calculations would require a tremendous effort. This effort itself must be considered as a tertiary disturbance, since the manpower involved would be withdrawn from other activities. Thus this way of tackling the problem might well introduce a diverging process.

A more practical approach looks more promising than this theoretical one. To diminish the number of factors to be taken into account some of them should be fixed in advance. If in the end the result should not be satisfactory, then these prefixed factors could be reconsidered. In literature it is common practice to make the first requirement the highest speed of control consistent with sufficient damping. It may be that in servomechanisms this approach is an excellent one, but it is rather doubtful whether in view of the various aspects dealt with in the foregoing, this is also true in process control. Generally speaking, it seems advisable in continuous processes to keep the flow rates as smooth as possible for all streams. Therefore it is more likely that process control would have to be as slow as compatible with sufficient deviation reduction. If control is slow, then the manipulation of the valve can be smooth and secondary disturbances will be small. Moreover, equipment can be simple, thus facilitating maintenance and adjustment. Settings can be conservative so that they will give satisfactory control under various load conditions. If then, under all conditions, deviation reduction is sufficient, nothing is left to be desired.

To illustrate this philosophy let us consider the attitude of a control man in an oil refinery with respect to flow, pressure, and temperature control. Generally speaking, flow and pressure control are considered as the easier ones, temperature control as the tougher one. Now the deviation-ratio curves for these types of control differ only in the realizable values of the resonance frequency. Usually flow and pressure respond much faster to valve manipulations than temperature does, resonance frequencies being accordingly higher. Since total plant control incorporates a great many controls with different speeds of response, the ultimate deviations in product quality are likely to be dependent mainly on the slower controls. The faster controls are good enough if there is sufficient difference between "slow" and "fast." This means, that zone 1 in Fig. 2 of the faster controls should cover amply the resonance zone of the slower ones. Much more is usually unnecessary, since more rapid fluctuations are likely to be averaged out in the more slowly responding parts of the plant. In many cases these requirements can be met easily. This is, for instance, the reason why slow mercury manometers often do a good job in flow control.

Various instrument manufacturers have realized that, in cases like this, simple controllers can be used, even with nonadjustable proportional band and reset rate, and there is a tendency to have these controllers acting with a very wide proportional band, so that they approximate pure proportional-speed floating action (9). This seems a sound approach, since this type of action gives very smooth valve manipulation. Of course such instruments are not as flexible as the more complicated three-term controllers. So if sufficient knowledge is lacking about the deviations to be expected and about the deviation reduction required, then to be on the safe side a three-term controller could be chosen.

In view of this philosophy, the question arises whether there is any use in determining exactly, by measurement or by other means, the dynamics of industrial processes. If these dynamics were known, then the deviation ratio could be determined exactly, provided also the controller dynamics were known. Although usually the knowledge of controller dynamics is taken for granted, determining them is not as simple as it looks (10, 11). But let it be assumed for the moment that all data are available. Then it is still of little use to know one factor accurately if other important factors are known only roughly.

Analyzing disturbances, their consequences, and the consequences of control is just as important as measuring process dynamics (12). Complete quantitative analysis being impracticable, commonsense judgment should include all the elements mentioned. An experienced control man in the plant has these things at the back of his mind, and so is able to find better settings by experience, than by the help of rules for optimum setting that are based on a thorough treatment of a small part of the entire problem. Of course there is no objection to having some accurate information on process dynamics. It even can be very helpful, in particular in research work, for finding out for which frequencies deviation reduction is possible, for which not, and how the resonance frequency can be shifted. However, it should be kept in mind that this type of information alone is not the key that unlocks the gate to the paradise of process control.

So far it has been tacitly assumed that what control does offers sufficient means of deviation reduction. However, also the plant contributes to deviation reduction, as was shown in the first section. There should be a reasonable balance between the measures taken in the plant for this purpose, and the effort put into automatic control. The more stable the plant is in itself, the more disturbances are eliminated beforehand and the more easily control can meet requirements. Quite naturally, violent disturbances in a resonance zone are very awkward, especially if successive stages of the plant have overlapping resonance zones. If it is impossible to detune them, then interstage storage may well prove worth while.

How Automatic Control Works

In this section some simple graphical relationships between the deviation-ratio curve and the Nyquist diagram will be considered. These relationships will be illustrated by characteristic examples (Figs. 6, 8, and 9).

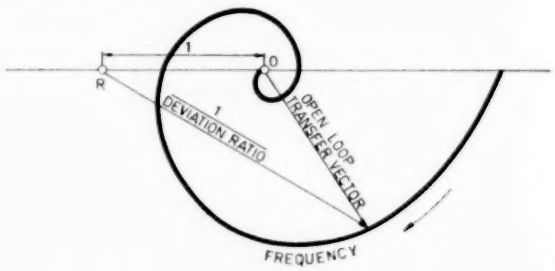


Fig. 5 Deviation Ratio and Nyquist Diagram

Fig. 5 shows how for any frequency the deviation ratio can be found from the Nyquist diagram (6).

Fig. 6 shows how the proportional-band setting affects the deviation ratio. The change in the proportional band is represented in the diagram by the change in the unit of length (*RO*). Thus when the proportional band varies from infinity to the limit of stability, point *R* moves from $-\infty$ to *A*. Accordingly, point *B*, which corresponds to the peak in the deviation-ratio curve, moves from *C* to *A*. The figure shows also the movement of points *D* and *E*, for which the deviation ratio equals unity. These points are related to the boundaries between the three

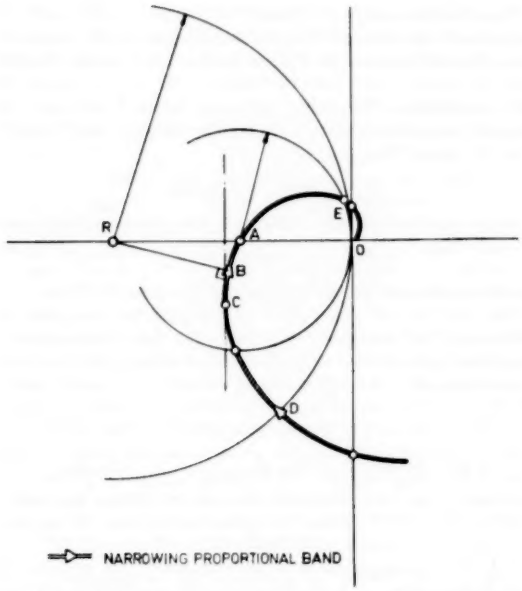


FIG. 6 EFFECT OF PROPORTIONAL BAND ON DEVIATION RATIO

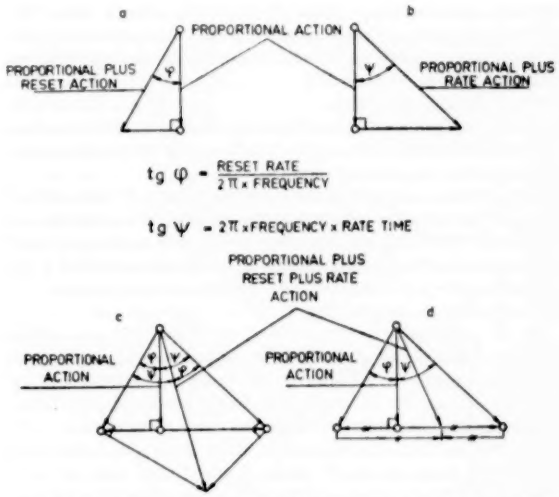


FIG. 7 CONSTRUCTING NYQUIST VECTOR FOR MULTIPLE ACTIONS FROM ONE FOR PROPORTIONAL ACTION ONLY

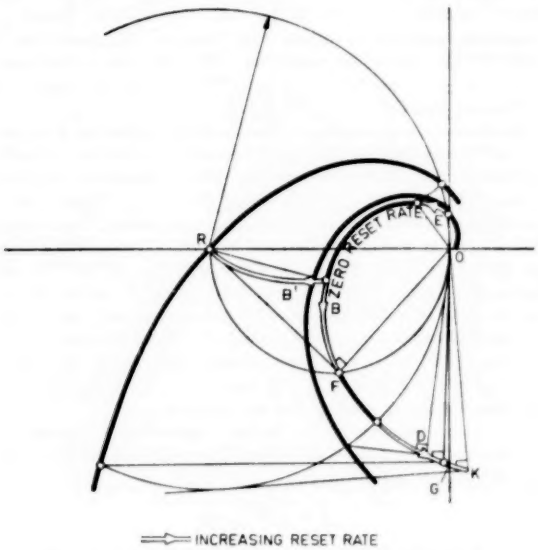


FIG. 8 EFFECT OF RESET RATE ON DEVIATION RATIO

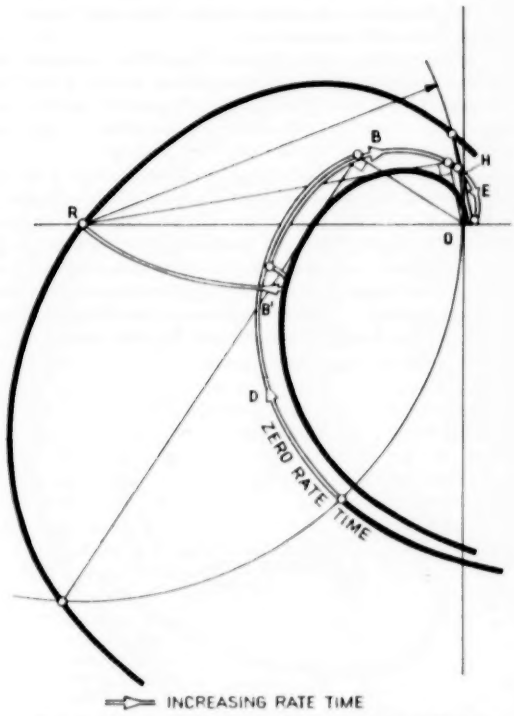


FIG. 9 EFFECT OF RATE TIME ON DEVIATION RATIO

zones. It is not proposed here to identify the points and boundaries. In Fig. 2, for instance, point E would be hard to find. The movements of D and E, however, are similar to those of the boundaries.

Fig. 6 explains Fig. 3(a).

Adding reset action to proportional action changes the shape of the Nyquist diagram. For any frequency the open-loop vector for proportional plus reset action can be found by simple construction from the vector for proportional action, as is shown in Fig. 7(a).

Fig. 8 shows how the deviation ratio changes if for constant proportional band the reset rate increases from zero to the limit of stability. In this figure three Nyquist plots are drawn, one for zero reset rate (only proportional action), one for the limit of stability (passing through point R), and one for some intermediate value of reset rate. The movement of point B shows how the frequency of the peak of the deviation-ratio curve changes. Point B' represents the corresponding travel of the peak value. Points E and D represent frequencies of unity deviation ratio. It is interesting to note that the movement of point D reverses its direction at K. The frequency at which the system starts

hunting, represented by point F, can be decreased further by widening the proportional band. In the extreme case of infinitely wide proportional band (proportional-speed floating action) point F reaches point G.

Fig. 8 explains Fig. 3(b).

For the addition of rate action to proportional action one can use a construction similar to that for the addition of reset action, as is shown in Fig. 7(b).

Fig. 9 shows how the deviation ratio changes if for constant proportional band the rate time increases from zero to the limit of stability. This figure is so like Fig. 8 that it does not need much explanation. The hunting frequency will increase further if the proportional band is widened. Its extreme value corresponds to point H. Fig. 9 explains Fig. 3(c).

If the controller has proportional plus reset plus rate action, then the construction of the Nyquist diagram depends upon the type of interaction in the controller. If the controller consists of the cascade connection of a proportional plus reset unit and a proportional plus rate unit, then the two constructions given before can be applied in succession, the order of succession being immaterial, as is illustrated in Fig. 7(c). If the controller is of the pure three-term type, then the construction is somewhat different, as is shown in Fig. 7(d). For other controller transfer functions the constructions are different, but for a mere qualitative study of deviation-ratio behavior the differences between the constructions do not much matter.

A diagram similar to the Nyquist diagram can be drawn for feedforward control. The deviation without control comes from the measured variable. The manipulated variable gives a correction. The deviation with control is the difference between the deviation without control and the correction. The corresponding vector represents also the deviation ratio if the deviation without control is put equal to unity. The solid line in Fig. 10 represents the case where the controlled variable responds much more slowly to the manipulated variable than to the measured variable. Accordingly, with increasing frequency the correction lags behind and decreases with respect to the deviation without control. Here the behavior of the deviation ratio is similar to that in feedback control, and the three-zone picture is quite adequate.

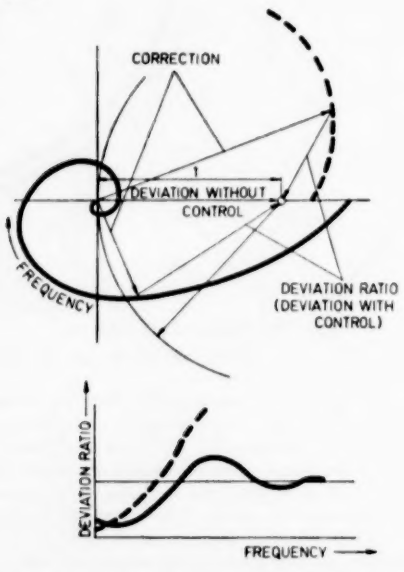


Fig. 10 Deviation Ratio for Feedforward Control

The dash line in Fig. 10 represents the case where conditions are just the reverse. With increasing frequency the correction has a phase lead here and is increasing relative to the deviation without control. So there is considerable overcorrection for high frequencies. By adding time lags in the faster path, for instance, between measuring element and valve, improvements could be made (8).

Westcott's Theorem

Westcott's theorem when applied to the deviation ratio reads

$$\overline{\log \left| \text{ deviation ratio } \right|} = 0$$

the average taken over the frequency from zero to infinity.

This theorem has been proved for feedback control, under the assumption that the open-loop transfer function approaches zero for increasing frequency at least as fast as the inverse of the frequency squared. It is quite easy to construct on paper a control system that does not satisfy the foregoing condition. A simple example is a first-order process controlled by a proportional controller. However, a first-order process does not exist in reality and neither does a pure proportional controller without any time lags. All physical systems are of a very high order, even of infinitely high order owing to distributed time lags. So in practice Westcott's theorem always will hold.

Now the objection could be reformulated. In practice the plant could have one major time constant, so that for all practical purposes its behavior can be represented by a first-order differential equation. Does Westcott's theorem then make any sense? It does, since even in this case there will be a resonance frequency, although it will be very high, its value being determined by the minor time constants. Also, the peak in the deviation-ratio curve can be very low, but that is not necessarily so. It depends upon the controller setting.

Westcott's theorem sets no limit to the speeding up of control by introducing higher and higher derivatives into the controller equation. This would shift the resonance frequency to higher values. At the same time it would give rise to more violent movements of the control valve, and here are practical limitations. In this connection reference may be made to Westcott's companion paper (13). However, there seem to be peculiar differences between servomechanisms and process control in this respect. In servomechanisms the power delivered can be set proportional to the square of the signal, whereas in temperature control, for instance, there is a constant heat supply in the absence of any disturbance and, if they occur, the additional heat supplied is directly proportional to the signal.

Westcott's theorem is not directly applicable to feedforward control. In Fig. 4 the relationship between deviation x and initial disturbance y is given by

$$x = y\,(P - Q)$$

where P and Q are the transfer functions along the paths indicated in Fig. 4.

Without control we have

$$x = yP$$

Consequently, the deviation ratio equals $1 - Q/P$.

This expression is of the same form as the inverse of the deviation ratio for feedback control, which reads $(1 + T)$, where T is the open-loop transfer function. The important point now is to know in what respects Q/P behaves differently from T.

Without detailed information about the plant concerned, there is no reason to suppose that, in general, with increasing frequency Q/P would approach zero at least as fast as the inverse of the frequency squared. There are certainly many practical cases where the condition is satisfied, for instance if the time lags in

the plant, Fig. 4, are equal. Then Q comprises in addition the time lags of the measuring means, the controller, and the manipulating means. This situation is not unusual in ratio flow control.

Let us now consider the possible singularities of the function $1 - Q/P$. It can have both poles and zeros in the right half of the complex frequency plane. $1 + T$ has no zeros there, if control is stable. Another reasoning due to Westcott can be followed now.[3] Consider a zero of $1 - Q/P$ in the right half plane. Replacing this zero by its image with respect to the imaginary axes will not change the amplitude-frequency relationship of $1 - Q/P$. But the theorem is correct if the singularities lie in the left half plane. Consequently, zeros and poles in the right half plane do not affect the validity of the theorem.

Thus a sufficient condition for Westcott's theorem to hold also for feedforward control is that the path containing the controlling means is at least two orders slower than the direct path.

BIBLIOGRAPHY

1 "Optimum Settings for Automatic Controllers," by J. H Ziegler and N. B. Nichols, Trans. ASME, vol. 64, 1942, pp. 759–768.

2 "The Practical Application of Frequency Response Analysis to Automatic Process Control," by C. I. Rutherford, Proceedings of The Institution of Mechanical Engineers, vol. 162, 1950, pp. 334–354.

3 "The Influence of Measuring and Transmission Lags," by A. R. Aikman, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, pp. 205–216.

4 "Automatic Feedback Control," by W. R. Ahrendt and J. F. Taplin, McGraw-Hill Book Company, Inc., New York, N. Y., first edition, 1951.

5 "Automatic Control Terminology," Terminology Committee of the Industrial Instruments and Regulators Division, ASME Paper No. 52—SA–29 (not published).

6 Loc. cit., by W. R. Ahrendt and J. F. Taplin, p. 91.

7 "The Development of Relationships Concerning the Frequency Bandwidth and the Mean Square Error of Servosystems From Properties of Gain-Frequency Characteristics," by J. H. Westcott, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, pp. 45–64.

8 "Some Aspects of Cascade and Multi-Element Control Systems," by J. C. Farquhar, Proceedings of the Third World Petroleum Congress, Section W l, E. J. Brill, Leiden, Holland, 1951, pp. 493–505.

9 "Some Recent Developments in Process Control," by W. H. Howe, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, pp. 217–224.

10 "The Characteristics of Air-Operated Controllers," by A. R. Aikman and C. I. Rutherford, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, pp. 175–187.

11 "Analysis of Pneumatic Controllers," by J. M. L. Janssen, *Automatic and Manual Control*, Butterworths Scientific Publications, London, England, 1952, pp. 189–200.

12 "The Automatic Controller's Organic Function in the Plant," by J. M. L. Janssen, *De Ingenieur*, vol. 64, 1952, pp. O125–O131.

13 "The Synthesis of Optimum Feedback Systems Satisfying a Power Limitation," by J. H. Westcott, published in this issue, pp. 1253–1263.

Discussion

J. H. Westcott.[4] The author performs a public service in his introduction. His words should be framed and hung in a prominent position in every engineering office where frequency-response ideas are beginning to seep in.

This story of frequency response really began with the alternating-current vector diagrams with which every electrical engineer is primed from an early age, but the story has passed on from single-frequency vector diagrams to vector loci, being a curve traced by the tip of a vector as frequency varies. Finally, the picture is completed and needs no further elaboration, when the concept of frequency is extended to include complex values of the frequency variable. Thus we come the full journey to functions of a complex variable and here, in the theory built around these functions, reside all the answers to linear-circuit problems.

Effective use has been made in the paper of a theorem the author attributes to the writer. The writer would like to fill in the background. The theorem appears in a paper presented at the Cranfield Conference in 1951. Janssen took an active part in the discussion of the paper and has pointed out that the theorem applies directly to deviation ratio. The theorem follows directly from the classical theory of functions of a complex variable which was invented, if that is not too strong a word, by the French mathematician, Cauchy, in the early part of the nineteenth century.

At that time practical applications were not in mind and it is only recently that it has been relevant to point them out, the most famous of these occasions being Harry Nyquist's development of the stability criterion. Also high on this list is the work of that master of function theory as applied to networks, Hendrik Bode, to whom the writer is personally indebted for having interested him in network problems. There is no better introduction to this fascinating field than his book "Network Analysis and Feedback Amplifier Design."

A fact which emerges most clearly from the present paper is that, abstruse as these theorems in function theory may seem, it is still possible to give them a simple geometric or graphical interpretation which appeals directly to the imagination of the engineer and gives him a sound philosophy in terms of which to frame his design decisions.

R. H. Macmillan.[5] If the transfer functions of the process and the control are, respectively, $Y_p(\omega)$ and $Y_c(\omega)$, then the "deviation ratio," as defined by the author in the second section of his paper, has the value $1/(1 + Y_pY_c)$. This is true whether the disturbance is at the set point or acts directly on the load. By plotting the Nyquist locus, Y_pY_c, for the system, the reciprocal of the deviation ratio can be read directly as the vector from the $(-1,0)$ point. Variation in controller gain is easily accounted for by a shift of this point along the negative real axis.

Now, for any given process, the locus of Y_p for varying frequencies can be determined experimentally; or, with simplifying assumptions, it may be plotted from an analytical approximation to the process-transfer function. This harmonic-response locus for the process cannot be altered appreciably at the designer's will, whereas the shape of that for the controller can be adjusted to some extent at his discretion. There seems, therefore, to be some advantage in separating the two loci when determining the quantity $(1 + Y_pY_c)$.

When the controller setting is such that sustained oscillation can occur for zero input, the deviation ratio is infinite and its reciprocal is zero, so that

$$1 + Y_pY_c = 0$$

which can be expressed as

$$-\frac{1}{Y_p} = Y_c$$

The locus $-1/Y_p$ has been called the "potential correction" curve and would appear somewhat as in Fig. 11 of this discussion, for a typical stable process, with frequency ω_p graduated along it and increasing in the direction shown. The locus Y_c for a simple proportional controller involving a single exponential lag time is a

[3] Bibliography (7), p. 64.

[4] Lecturer, Imperial College of Science and Technology, University of London, London, England.

[5] University Lecturer, Cambridge University, Cambridge, England.

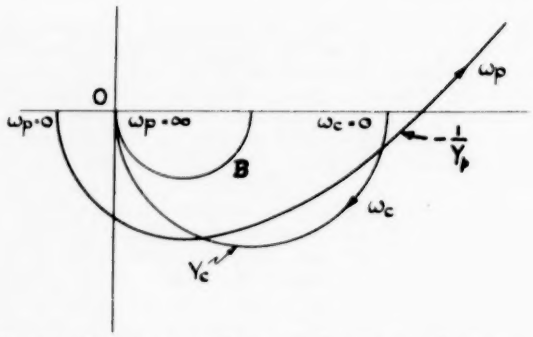


FIG. 11 CHARACTERISTIC CURVES OF PROCESS AND CONTROLLER

semicircle as shown, with frequency ω_c increasing toward the origin; reducing the time lag leaves the shape of the locus unchanged but alters the frequency graduations, shifting them away from the origin. Altering the gain (proportional band) of the control changes the size of the locus in direct proportion; locus B is drawn, for example, for half the controller gain. The method outlined by the author in his Fig. 7 is suitable for obtaining the Y_c locus.

Clearly now, if these two curves cut, and the frequency graduation is the same on both curves at the intersection, then the condition for sustained oscillation is satisfied and it will occur at that frequency. For the particular gain indicated in the figure, the curves intersect twice. There are thus two values of controller time lag for which the system will be on the limit of stability. If the gain is reduced, a stage is reached where the curves no longer intersect. At this controller setting instability cannot occur, no matter what is the controller time lag. In this way a curve showing the limit of stability for various controller settings can be obtained as in Fig. 12, herewith.

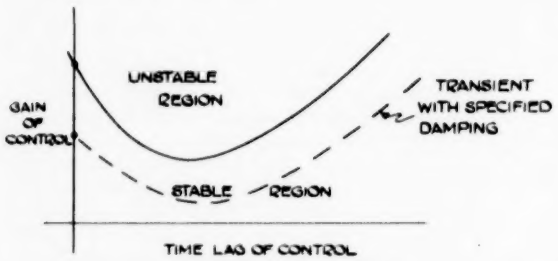


FIG. 12 STABILITY REGIONS

In practice, the designer requires a system which is suitably damped, not one on the limit of stability, i.e., with zero damping. The method outlined can be extended readily to obtain a chart similar to Fig. 12 for a system with specified damping. By using the conformal plotting technique proposed by Campbell[6] and Profos,[7] one obtains for the process the characteristic of Fig. 13 of this discussion, and by modification of its method of construction one also obtains the new control characteristic shown. From the intersections of these curves we can obtain the broken line in Fig. 12.

The actual construction of charts such as in Fig. 12 would be

[6] "Use of Nyquist Diagram," by W. W. Campbell, Min. of Supply, Servo Library, Reference G. 74, 1945.

[7] "A New Method for the Treatment of Regulation Problems," by P. Profos, *Sulzer Technical Review*, No. 2, 1945.

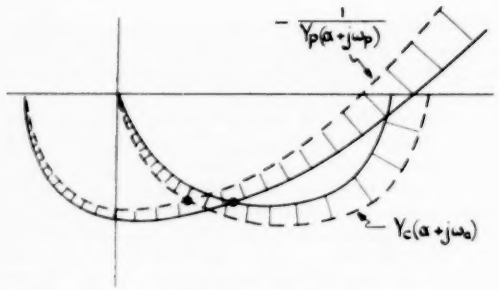


FIG. 13 MODIFIED CHARACTERISTIC CURVES

facilitated greatly by using logarithmic plots instead of the polar form, as changes of gain and time lag can be accomplished by simply shifting the curves vertically or horizontally. The polar form is perhaps clearer for purposes of explanation.

J. E. BIGELOW.[8] The author has made a valuable contribution to the development of a philosophy of control which is conformable to both the "theorist" and the "practical men." In the first instance, the deviation ratio which he explains is determined from frequency-response information by elementary automatic-control theory. In the second, it gives a picture of control quality in simple language. A third important advantage in the use of the deviation ratio is that it applies as well to feedforward as to feedback control.

As the author points out, however, the emphasis in his paper is on what the control does apart from what the plant does. He recognizes, of course, that analysis of disturbances and its consequences is as important as measuring process dynamics but assigns this task to common-sense judgment which makes quality of control with a given system an important function of the experience of the control man. For a quantitative measure of overall control quality, one may feel forced to turn back to the "deviation reduction factor." It might be more in line with the desired ends of his control philosophy to carry further the presentation of information as a frequency function with disturbance data converted to this form. While such a project might be beyond the scope of the paper, it would seem more fitting to recognize the problem in this light rather than to go back to the common sense of experience as the only practical answer.

The section discussing how automatic control works would have been simpler if a Bode diagram had been inserted as a step in the development of the Nyquist diagram. Most readers will not be troubled by this. However, the section was included to furnish a theoretical background, and those who had not previously acquired it might be troubled.

The points of contention in the foregoing are minor in importance compared with the main point of the paper. The philosophy developed spawned a new term of possibly great usefulness and lays a foundation of logic upon which to proceed.

SIDNEY LEES.[9] The author is to be congratulated for his efforts to bring theorists and their theories to the practical men. He should not be allowed to remain alone in his efforts. It seems worth while for engineers to re-examine the relationship between theory and practice at least once a year, like visiting the dentist. In his introduction the author lists some of the grievances which practical men hold against the theorists.

[8] Engineer, General Electric Company, Schenectady, N. Y.

[9] Assistant Professor of Aeronautical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

The body of his paper attempts to explain the frequency-response approach to control-system design. He has made a modest attempt to bridge the gap he describes. The first part of his paper appeals to the discusser since it is similar to the methods the latter has used to describe simple automatic-control-system design theory to students.

On the other hand, the jump from the deviation-ratio concept to Nyquist's criterion is too abrupt for a novice in the frequency-response method of design. The discussion of Westcott's theorem becomes rather esoteric in the light of the objectives stated in the introduction. One suspects that the author had more in mind when he planned the paper. Perhaps he will enlarge upon these ideas in the near future.

The deficiencies of the theoretical approach in system design described in the introduction are worthy of more discussion. What is the purpose of theory in engineering design? It would seem that the answer lies in the value and merit of an engineer. The "value" of an engineer may be described in simple terms. It is his ability to produce physical equipment which satisfies given specifications and requirements. An implication of this description is the ability to discover the possibility of meeting given specifications. The engineer should be able to disclose the existence of nonfeasible or conflicting specifications. It is important that he should be able to tell when a given scheme will not work.

The "merit" of an engineer is measured by the speed with which he produces the required equipment after being given the specifications. Generally, the early history of engineering is a description of trial-and-error methods. In many situations it is still the easiest and fastest method.

Difficulties are encountered in standard situations when the specifications and requirements are unusual. For these applications one resorts to theory to demonstrate that it is feasible and practical to invest money and effort in producing the desired equipment. Theory is also useful when new ground is being broken and there is no past experience upon which to draw. Some guides and signposts are required to tell the designer when he is on the main road and when he is in the swamp.

Theory is most useful to students and young engineers who do not have much experience as a guide during the trial-and-error stages. For these people it is necessary to organize known experience so that they can become effective as quickly as possible. The theory should be so well developed that it can be applied with a minimum of effort. The measure of a successful engineering theory is not how completely it describes a complicated situation but how simply it can be expressed and still account for enough of the facts to carry out a practical design. The deviation-ratio curve in Fig. 2 of the paper and the accompanying text is an example of the simple but effective approach.

The danger always exists that the tyro is so beguiled by the closed forms of theoretical solutions he loses contact with the problems encountered in wrestling with the perversities of inanimate matter. To quote the delightful phrase of Prof. Arnold Tustin, "He is transported to that curious world of decibels and negative frequencies where filter experts live."[10] In addition to the conditions mentioned previously, a successful engineering theory also should include a description of the pitfalls awaiting him who follows too closely. The tyro must be encouraged to use imagination and intuition tempered by judgment.

The author points out in his introduction that theory is an abstraction and a simplification of reality. This is most emphatically true. It always should be kept sharply and clearly in view. The theory applied in a given situation should produce numerical results as quickly as possible so that the designer may acquire the proper components in the most likely combination and get on with the job of making the system work. It is a truism but nonetheless a most important condition of design that the real problems are encountered only when the physical equipment is under construction. In the final analysis a system design is completed with a soldering iron, a screw driver, and a wrench. The further along the design proceeds, the cruder and simpler is the theory. The trial-and-error stages become more significant.

In summary then, it appears that the difficulties any practitioner encounters, according to the author, are due partly to the novelty of the theories but more because the theories are incompletely digested. It is necessary to have a firm grasp of the nature of the physical processes partaking in the system over and above the mathematical techniques required to solve differential equations. Solutions in physical terms, including the limitations that exist because real physical equipment is involved, can shorten the mathematical labor. Engineering judgment and intuition must be employed to bring the design into practical and useful form. The direction of theory for practicing engineers should be to reduce the mathematics to a minimum and not to the production of elegant solutions. The proof of the theory is always the production of real working equipment.

[10] "The Nature of the Load as a Factor in the Design of Electrical Servos," by A. Tustin, Trans. Society of Instrument Technology, vol. 4, March, 1952, p. 13.

Author's Closure

The introduction to the paper has given rise to extensive comments by J. H. Westcott and Sidney Lees. The author is very grateful indeed to them for enlarging upon the subject.

Westcott is certainly too modest if he transfers the credit for finding his theorem to Cauchy. Since Nyquist's development of the stability criterion nineteen years have elapsed before Westcott made the theorem "follow directly from the classical theory of functions of a complex variable." Apparently the theorem was so obvious that it was easy to overlook it.

The author does not object to framing his words as Westcott suggests. However, citing Lees with a slight modification, he should not be allowed to remain alone in his frame. First of all, he would like the company of his colleague E. Wieringa, who contributed essentially to the composition of the introduction; both would be very pleased by Lees' joining the party. His statement concerning the measure of a successful engineering theory is for the author a source of daily meditation. As a matter of fact, the text hangs in his office now, together with a complementary one due to J. von Neumann and O. Morgenstern: "There is no point in using exact methods where there is no clarity in the concepts and the issues to which they are to be applied."[11] Were it not for the limited size of his office, the author would like to give more of Lees' wisecracks a place on its walls.

Turning to the body of the paper, the author entirely agrees with Lees that in the paper the jump from the deviation-ratio concept to Nyquist's criterion is too abrupt for a novice in the frequency-response method of design. Accordingly, in the training courses given in the author's department this hurdle is taken much more smoothly. However, the paper was not written for the novice in the field. The author, rather, had in mind showing the initiates how the gap between theory and practice can be bridged, and since many belonging to this category know better how control works than what it actually does, he judged that the "how-it-works" part could be comparatively condensed. Consequently, for the relationship between the deviation-ratio and the open-loop transfer function, Fig. 5, reference was made to literature. For the same reason no need was felt for inserting a Bode

[11] "Theory of Games and Economic Behavior," by J. von Neumann and O. Morgenstern, Princeton University Press, Princeton, N. J., 1947, p. 4.

diagram in the development of the Nyquist diagram, as J. E. Bigelow suggested. Moreover, polar plots are easy to understand without any reference to either Bode diagrams or complex calculus. Thus the novice in the field should be able without undue effort to digest the graphical treatment of how control works when it is offered to him piecemeal in several appetizing courses. On the other hand, the discussion of Westcott's theorem is indeed only for those who feel at home in the accidented terrain of zeros and poles.

J. E. Bigelow does not like the author's turning back to common sense. Perhaps he will be partly satisfied by Fig. 14, with modified text taken from an earlier paper of the author (12). The

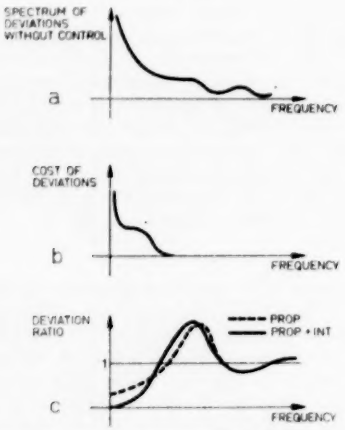


FIG. 14 MAIN FACTORS DETERMINING CONTROL QUALITY

author agrees that it is in principle possible to present information about disturbances as a frequency function, Fig. 14(*a*). In practice, however, this is far from easy. But the main trouble arises when one tries to evaluate the effects of disturbances economically. Fig. 14(*b*) only served the purpose of qualitatively demonstrating the importance of economic factors in the criterion for optimum control. However, at present the author really does not know how, on the basis of Fig. 14, a quantitative optimum-control criterion superior to common-sense judgment could be formulated.

R. H. Macmillan gives another example of Westcott's statement, that it is possible to give a simple geometric or graphical interpretation of what might seem abstruse theorems. In particular, the separation of the two loci of process and controller can be very useful indeed.

In order to avoid misunderstanding, the author would like to point out that Macmillan's controller transfer function is quite different from the one used in the paper. The behavior of a controller with pure proportional plus reset plus rate action, Fig. 7(*d*), is to be represented in Fig. 11 by a straight line perpendicular to the horizontal axis. Then it appears from Fig. 11 that for every value of controller proportional gain the system can be unstable for an infinity of combinations of reset rate and rate time. Moreover, the frequency of sustained oscillation is, within limits, independent of the particular settings of reset rate and rate time! Incidentally, the same conclusion can be drawn from Fig. 8 of the paper.

Macmillan's approach has also been used by O. Schäfer. Figures similar to Fig. 11 and Fig. 13 are found in his book.[12] The author wishes to mention this, not only in order to give credit to Schäfer, who already in 1950 demonstrated this method in a lecture in Frankfurt,[13] but also to show how difficult it is to keep abreast of literature. The author himself learned only recently that already in 1948 G. A. Philbrick[14] had defined a "regulatory transmission" which is identical with the deviation ratio.

The author would also like to give credit to J. Farquhar, who preceded him in advocating smooth and slow control. "With certain exceptions, such as cracking-plant tube stills, it will be found that the usual tendency is to adjust the controller sensitivity settings too tightly (i. e. to give too little damping) in an attempt to get unnecessarily 'straight line' control."[15] Unfortunately, Farquhar's paper was published after the author's paper had been submitted, so that no reference could be made to it in the text.

---

[12] "Grundlagen der selbsttätigen Regelung," by Otto Schäfer, Franzis-Verlag, Munich, Germany, 1953, pp. 113 and 121.

[13] "Stabilitätsuntersuchungen an Regelkreisen mit Methoden der Schwachstromtechnik," by Prof. Dr. O. Schäfer, Hauptversammlung des VDI, 5–8 September, 1950, Frankfurt-on-Main, Germany.

[14] "Unified Symbolism for Regulatory Controls," by G. A. Philbrick, Trans. ASME, vol. 69, 1947, pp. 47–67.

[15] "Plant Instrumentation," by J. C. Farquhar, *Reviews of Petroleum Technology*, vol. 13, 1951, pp. 86–98. Published by the Institute of Petroleum.

Frequency-Response Analysis and Controllability of a Chemical Plant

By A. R. AIKMAN,[1] LONDON, ENGLAND

The practical frequency-response analysis of two types of chemical-plant units, and the adjustment of the automatic controllers, are described. Available information on the magnitude, form, and location of disturbances is used to illustrate the effect of these factors on the quality of control actually obtained, and to emphasize that these factors are at least as important as the over-all characteristics of the plant. Even when all information on the disturbances is available, their effect on the quality of automatic control is not always easy to calculate, especially if the system has distributed parameters, but a description is given of an approximate calculation method which yields results of the right order of magnitude in many practical cases.

Introduction

In the past few years, a number of papers have been published on the application of frequency-response methods to automatic process-control systems. One of the first was by Ivanoff (1),[2] followed later by Farrington (2), with descriptions of the method applied to specific problems. Later, Rutherford published a paper (3) which showed how the frequency-response methods used so successfully in servomechanism studies may be applied to any process-control problem. Rutherford's paper illustrated the power of these methods for attacking problems of analysis or synthesis and showed how they can be applied using entirely graphical methods. The process analyzer described by Aikman and Halsall (4, 5), which was designed as a "plug-in" unit applicable to any air-operated control system with the minimum of practical difficulty, was another stage in the attempt to make frequency-response analysis a tool for common use by the process-control engineer.

In 1943, Profos (6) had applied frequency-response methods to the automatic control of temperature in steam superheaters, but this work was not generally known until the D.S.I.R. Conference in 1951, when it was referred to by Takahashi.

Further papers on frequency response methods by St. Clair, Coombs, and Owens (7) and others have followed, and some plant test results have been published by Higgins and McKnight (8). Methods for calculating the frequency-response characteristics of some types of heat exchanger, with some practical confirmation by experimental tests, have been published by Takahashi (9).

When work on the application of frequency-response analysis was started by the Central Instrument Section of I.C.I., Ltd., the principal use of the analyzer was in obtaining the over-all characteristics of plants so that controllers could be selected and adjusted to give their best performance. At the same time, the characteristics of the common types of air-operated controllers were studied and some results were published by Aikman and Rutherford (10) and Janssen (11). It was not long before sufficient confidence was gained to enable controllers to be adjusted without analyzing the plant, with greater certainty and speed than was possible by using former empirical methods. Analysis of plants under operating conditions remains a valuable aid to an accurate selection of the type of controller and calculation of optimum settings, but the main emphasis is now placed on the analysis of plants with the object of obtaining their dynamic characteristics. The purposes of this work are: (*a*) the correlation of the dynamic characteristics with controllability, and (*b*) the accumulation of information as a basis for accurate prediction of control-system performance at the design stage.

A large number of papers have been published in the past 20 years which describe the over-all characteristics of various theoretical plants, and the means of calculating stability, but it seems that attention has seldom been directed to the influence of the location of the disturbances in the control loop. This factor may be the most important in determining the controllability of the system. In this paper the practical frequency-response analysis of the control systems in two chemical-plant units will be described, and the results will be correlated, in so far as possible, with the information available on the disturbances entering the systems, and with the actual performance obtained.

Terminology

Terminology offers considerable difficulty at the present time, because many terms which are useful in describing the phenomena of frequency response are not included in the glossaries published by The American Society of Mechanical Engineers and the British Standards Institution. Both glossaries are being revised at the time of writing. It has been decided to use the terms defined in British Standard 1523:1949, with additional terms originally proposed by Rutherford (3) and defined in the Appendix. A recent paper by Young (12) which compares ASME, BS, and European glossaries of terms may be of interest in this connection.

The Automatic Control of a Spray Drier

Description of System. The material passing through the spray drier is an emulsion. The system is shown diagrammatically in Fig. 1. At intervals, the emulsion is moved up into the head tank from a lower floor; it cannot be pumped mechanically or otherwise subjected to high shearing rates because of the risk of coagulation. The physicochemical properties which determine the quality of the product are critically dependent on the drying temperature. The control system used had been chosen on the basis of experience as the one most likely to fulfill the requirements. The temperature-detecting element (originally a vapor-pressure bulb) was located in the offtake duct near its junction with the drying drum. The flow of emulsion is the regulated condition. Alternated filters remove larger coagulated lumps before they reach the regulating unit. The fan is driven by an induction motor, and the steam rate to the air heater is set manually. The performance of the system, as indicated by product quality, was not always satisfactory, and this was attributed to excessive temperature deviations. It was therefore decided

---

[1] Control Engineer, Central Instrument Section, Imperial Chemical Industries, Ltd.

[2] Numbers in parentheses refer to Bibliography at end of paper.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, July 1, 1953. Paper No. 53—A-22.

to investigate the system by frequency-response methods.

Practical Frequency-Response Analysis. In order to carry out the analysis a manifold was inserted into the air line between the controller and the valve during a shutdown period, so that the analyzer could be connected, and the change could be made from controlling to analyzing conditions, without upsetting the process operation. The analysis was carried out by generating a continuous sinusoid in air pressure and applying it to the valve motor. The "loop" was open while the analysis was carried out and the controller output (with controller on proportional action) was recorded by the analyzer, on the same chart which records the analyzer output. The recorded sinusoids on the chart are 4 in. long, hence the quantities required, amplitude ratio and phase lag, can be read off with ease at a number of different periods of oscillation. The amplitude of oscillation was as small as possible, to avoid nonlinearities and process interference. The analysis was concluded by applying successively two steady pressures to the valve motor and observing the corresponding steady values of plant output; i.e., the potential values under normal operating conditions. For the purpose of calculating controller adjustments, it is not necessary to know the potential values, but they are always determined when possible, in order to express the plant characteristic as a dimensionless quantity, attenuation. This is a particular requirement for fundamental interpretation of the data, as the attenuation is independent of such scalar quantities as valve-characteristic slope, measuring range, and proportional band. Full details of the practical procedure for carrying out a frequency-response analysis, and of the process analyzer itself, have been published (4, 5).

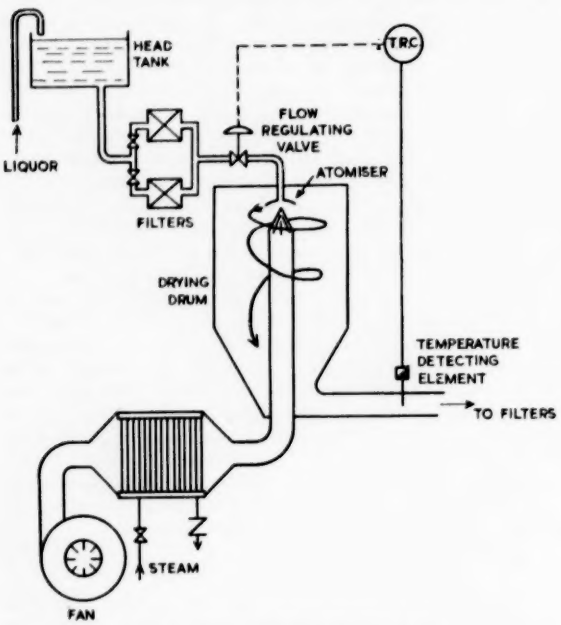


FIG. 1 SPRAY-DRIER LINE DIAGRAM

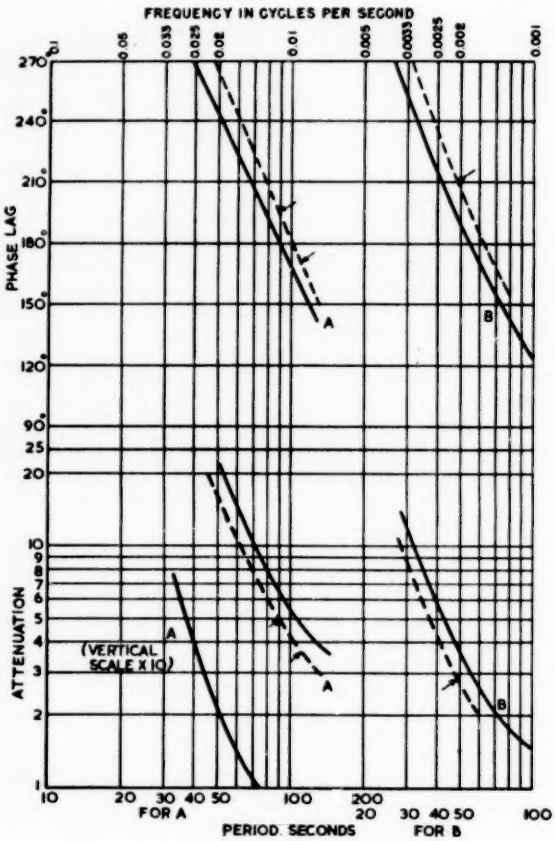


FIG. 2 FREQUENCY RESPONSE OF SPRAY DRIER

Controller Adjustment. The analysis was carried out with continuous oscillations; when the plant is controlled automatically subsiding oscillations are desired, and therefore to calculate the appropriate controller settings the curves must be modified. It was decided to adjust the controller for a subsidence ratio of e:1. The procedure for obtaining attenuation and phase curves for a given subsidence ratio from frequency-response-analysis results will be presented in full in another paper (13); the procedure for e:1 subsidence ratio is summarized in the Appendix, and the modified curves are dotted in Fig. 2.

The calculation of controller settings and the assessment of the value of different control actions can be carried out easily with this information available, and with a knowledge of the controller characteristics. For this application, integral action was necessary because the changes of level in the tank, which would have been sufficient to change the flow rate by a factor of 1.6 at a fixed valve setting, would have caused an unacceptably large offset. When using proportional-plus-integral control, an integral action time equal to the operating period is usually suitable. A longer integral action time unduly delays the elimination of offset, and a shorter requires a wider proportional band and causes larger deviations as a consequence. This ratio of integral action time to operating period is a practical useful compromise and causes a phase change of 9 deg in the controller, additional to the 180 deg inherent in the proportional action. The phase lag in the plant is therefore 180 — 9 = 171 deg, and this fixes the operating period at 110 sec. The controller gain at this period, to give the required subsidence ratio is, from Fig. 2, curve A, 3.7. This is a dimensionless quantity and, with the integral-action time to period ratio of 1, is equal to the proportional-control factor. To convert the proportional-control factor to proportional band, the potential values of the plant must be taken into account. The potential value was 1.0 psi pressure change applied to the valve, equivalent to 21 C change in the controlled condition.

The proportional band required is calculated from the expression

$$\text{Band width} = \frac{\text{potential value}}{\text{proportional control factor}} \times 12 = 8\ \text{C}$$

where 12 is the full output pressure range (in psi) of the normal air-operated controller.

When selecting settings for a three-action controller, the characteristics of the type of controller used must be taken into account. The characteristics of three-action controllers are not all alike, indeed they differ considerably. The characteristics of the controller available at the time have been described (10), where it was referred to as Type V. It is recognized by its possession of two feedback bellows, one within the other. The optimum settings in this instance are, to give the narrowest effective proportional band, nominal integral action time equal to twice the nominal derivative action time, and the former equal to one-half the operating period. Under these conditions the phase advance in the controller is 15 deg. At a plant phase lag of 195 deg, corresponding to this controller phase advance, the controller gain required is 5.0. The controller gain in a three-action controller is the product of proportional-control factor (which depends on proportional band), the interaction factor (1.8 in this instance), and the additional gain due to derivative action (1.11).

Hence the proportional control factor = 5/1.11 = 4.5, and the proportional band width required = (21 × 12/4.5) × 1.8 = 101 C. This is the proportional band width to which the instrument would be adjusted. The effective band width is 101/1.8 = 56 C.

As measures of the effectiveness of derivative action, the decrease in proportional band width and reduction in operating period may be chosen. In this instance, the effective proportional band width could be reduced from 68 deg C to 56 deg C, and the operating period could be shortened from 110 to 90 seconds, by using derivative action. It was therefore decided to use proportional, integral, and derivative control actions.

Results Obtained: Influence of Detecting-Element Lag. The controller was put into action with settings adjusted to their optimum values. The period was 90 sec as expected, and the subsidence ratio following a major disturbance was satisfactory. The temperature was apparently kept within ± 1.5 C of the desired value, but the product quality was occasionally unsatisfactory; further investigation seemed necessary. A thermocouple was inserted into the air stream near the vapor-pressure bulb, and the temperature was recorded on a self-balancing potentiometer recorder-controller. The temperature deviations recorded were nearly 3 times larger than those recorded by the vapor-pressure instrument; an independent check of the time lag of the vapor-pressure bulb, under identical heat-transfer conditions, showed it to be a single transfer lag of time constant 60 sec, whereas that of the thermocouple was 5 sec. The original controller was disconnected, and the controller of the potentiometer recorder-controller was now put into action. The over-all frequency response of the system, with the thermocouple instead of the vapor-pressure bulb, is shown in Fig. 2, curves B.

Controller adjustments were calculated in the manner illustrated previously, taking into account the characteristics of the controller used; this is referred to as Type IV in reference (10). In this plant (and in many others) it gives optimum performance with nominal integral action time half the operating period, and nominal derivative action time half the integral. The phase advance is then 30 deg, and the period, Fig. 2 B, is 50 sec. The interaction factor is 1.5, and additional gain due to derivative action is 1.27. Hence, the proportional band required was (for a total controller gain of 2.9) 166 C.

The results obtained are shown in Fig. 3, which is a section of a typical chart of the potentiometer recorder-controller. Since the attenuation in the thermocouple at the operating period is only 1.1, it is safe to assume that the temperature in the plant is maintained within about ± 2.0 C of the desired value. The chart from the vapor-pressure bulb recorder, which was left in place,

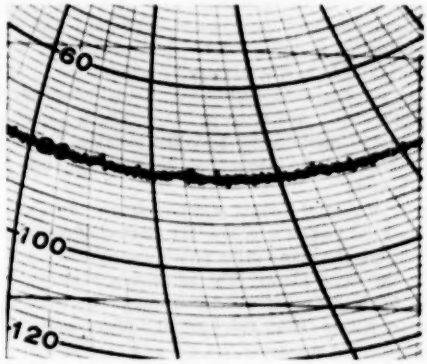


FIG. 3 RECORD OF SPRAY-DRIER EXIT TEMPERATURE USING THERMOCOUPLE RECORDER-CONTROLLER

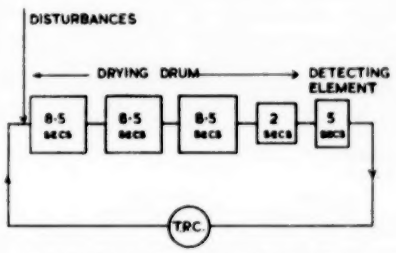


FIG. 4 BLOCK DIAGRAM OF SPRAY DRIER

showed deviations of about ± 0.3 C during the same period of time. Since the investigation was completed, product quality has been consistent and satisfactory.

Assessment of Spray-Drier Controllability. The frequency-response characteristics of the spray drier alone can be obtained by subtracting from the over-all frequency response the known frequency response of the detecting element, transmission pipe, and regulating unit. A laboratory test showed that the latter two items made no significant contribution at the periods of oscillation used. It was deduced from the frequency response that the system, exclusive of the detecting element, behaved as though it were composed of three transfer lags all equal to 8.5 sec, plus a distance-velocity lag of 2 sec. A block diagram of the control loop, with one block per lag, is given in Fig. 4. All possible disturbances, such as liquid head, filter pressure drop, partial obstruction of valve, and steam and air temperature, enter the control loop at one point; namely, just after the controller, and before the plant lags which are all associated with the drying drum. In their passage through the plant they will become attenuated before reaching the detecting element. It has been shown theoretically (14) that the deviation reduction factor for a step disturbance applied at any point in a control loop is, for a system with lumped parameters, closely related to the attenuation between disturbance and detecting element at the operating period and is approximately equal to it. The drying drum is strictly not a lumped-constant system, but one with distributed constants; but without greater knowledge of the flow pattern and the mechanism of heat and matter transfer, the analysis cannot be more exact, and the approximation must suffice for this examination.

When the Type V controller with the vapor-pressure bulb was in operation, the calculated attenuation in the drum at 90 sec period was 1.5. With the Type IV controller and the faster detecting element, although the controller gain was less, the attenuation in the drum rose to 3.1 as a result of reducing the operating period. Since measurements of potential value or exact form of disturbances could not be made, absolute values of deviation reduction factor could not be calculated. While the attenuation in the drum was doubled by a reduction in the operating period, the maximum deviations in actual temperature were reduced from about ± 4 C to ± 2.0 C, and it is evident that there is an approximate proportionality between the deviation reduction factor and the attenuation in the appropriate parts of the control system.

Alternative Method of Control of Spray Drier. It is of interest to assess the relative merits of another method of controlling the exit temperature of a spray drier, which was used at one time for controlling the same plant. The method was to regulate the steam flow to the air heaters. A reasonable figure for the total lag in the air heater (from valve to air temperature) would be about 200 sec. For simplicity, we will assume it to be two equal transfer lags of 100 sec each. An additional distance-velocity lag of about 3 sec must be added for the transport of the heated air from the heater to the top of the drum. The block diagram of this system is shown in Fig. 5. The operating period using the thermocouple and the Type IV three-action controller is calculated to be 230 sec. At this period, the attenuation in the drying drum, calculated from the known lags, is 1.03, and the effect of suddenly applied disturbances in liquor flow would be observed with their potential value almost undiminished. The effect of these disturbances is but little attenuated in the drum itself before being detected, while the potential correction applied by the controller has to traverse the large lags in the heater before becoming effective. Obviously, the flow of liquor would have to be controlled closely in order to obtain good quality of control with this system; further, it would require capital and maintenance expenditure on two controllers. Many of the materials which are processed in spray driers do not lend themselves well to flow control or positive injection, and the first system described is then the only one applicable, as experience has shown.

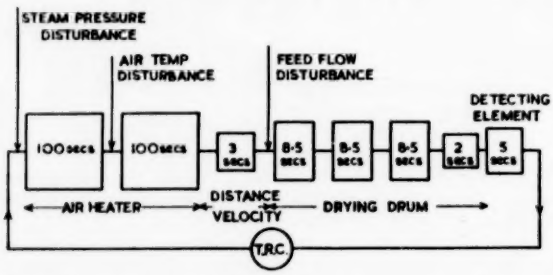


FIG. 5 BLOCK DIAGRAM OF SPRAY DRIER; ALTERNATIVE ARRANGEMENT

AUTOMATIC CONTROL OF A HEAT EXCHANGER

Description of System. In the example of the spray drier, the purpose was to illustrate the importance of the point of incidence of the disturbance, and of the reduction of amplitude of the disturbance which takes place in the plant before its effect is detected. In systems such as the spray drier, which can easily be reduced to detailed block diagrams, the points of incidence of known disturbances can easily be located, but this may be much more difficult in the case of indirect-contact heat exchangers, many types of which have distributed parameters.

Such a heat exchanger is shown in Fig. 6. This performs a critical duty in preheating an aqueous solution by steam to a desired value of temperature in the region of 110–120 C; the temperature is chosen according to the product required. The solution is pumped through the preheater into the reactors. The quality of the product depends on the quality of temperature control in the reactors, which in turn depends greatly on the temperature fluctuations at the preheater exit.

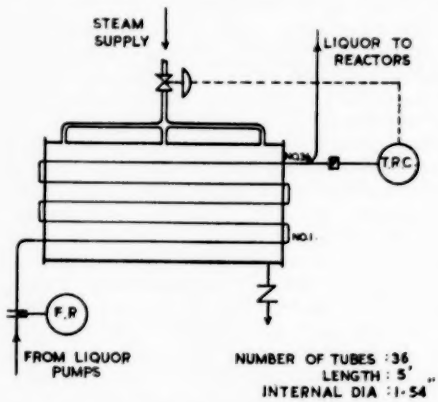


FIG. 6 HEAT-EXCHANGER LINE DIAGRAM

The principal disturbances likely to affect this control system are known to be changes in load and variations in steam-supply pressure. The investigation was requested because of the poor quality of temperature control which was being obtained.

Frequency-Response Analysis: Original System. The first analysis was carried out when the temperature-detecting element (inserted in the last preheater tube) was a sheathed vapor-pressure bulb. The analyzer output was connected to the steam valve, and the input was derived from the temperature controller. The analysis was carried out without difficulty under normal operating conditions, although the periods of oscillation required were of the order of 8–15 min. The controller gain required was quite low (5.0), which corresponded to a wide proportional band (75 deg) using all three control actions of a Type V controller. The predicted period of damped oscillation was 11 min. When this controller was put into action with the calculated settings, the expected period of oscillation and subsidence ratio were obtained, but the deviations were as much as ± 5 C, which did not satisfy the requirement for maximum deviation ± 1 C. It was suspected that the temperature-detecting element made a significant contribution to the over-all system response, and an independent test showed that the total lag of the element assembly, under equivalent heat-transfer conditions, was about 4 min, and further that this total was composed of two unequal transfer lags. Without attempting to dissect the over-all response into its component equivalent lags, it was decided to install a much faster temperature-detecting element and to investigate the plant more thoroughly after that had been done. The detecting element finally chosen was an iron-constantan thermocouple welded into the top of a sheath cut down to $^1/_{16}$-in. wall thickness. The time constant of the transfer lag of this assembly was 1 sec.

Frequency-Response Analysis: Modified System. A series of analyses were carried out to elucidate the characteristics of the heat exchanger itself. The first analysis was carried out from

steam valve to exit temperature at normal load, and the results are shown in Fig. 7, curves A. Correcting for the desired subsidence ratio, the controller gain required on proportional action is 60, which means a proportional band width of 3.0 C, since the potential values were 1 psi applied to valve equivalent to 15 C. The predicted period was 54 sec. The performance achieved on a 50 per cent load change, Fig. 8, with a steady steam-supply pressure is shown in Fig. 9; the deviation is about 0.6 C, accompanied by an offset of the same order of magnitude, since integral action was not used. Integral action was neither desired nor necessary under these conditions. If the steam valve had remained in the original position while the load change was applied, the exit temperature would have dropped 37 C. The deviation reduction factor is therefore about 60.

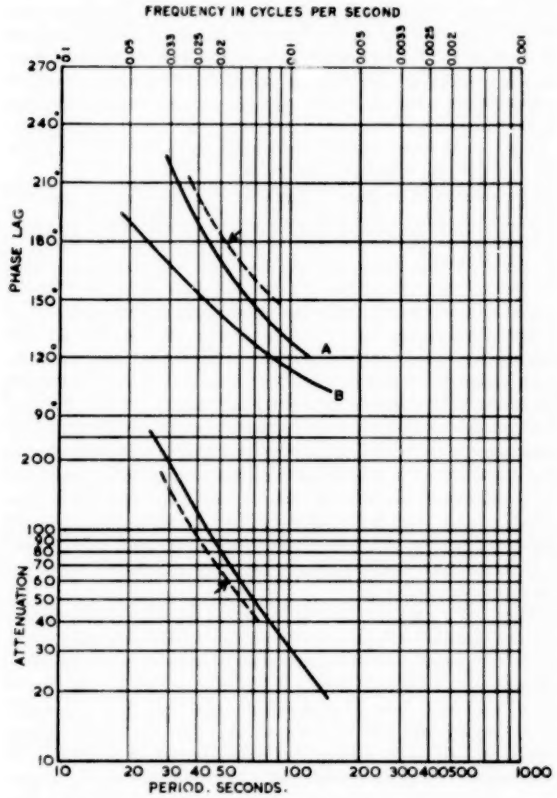


FIG. 7 FREQUENCY RESPONSE OF HEAT EXCHANGER

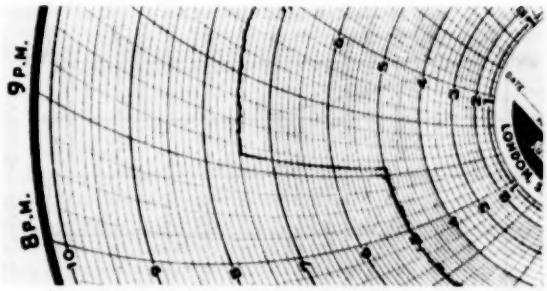


FIG. 8 LOAD CHANGE APPLIED TO HEAT EXCHANGER

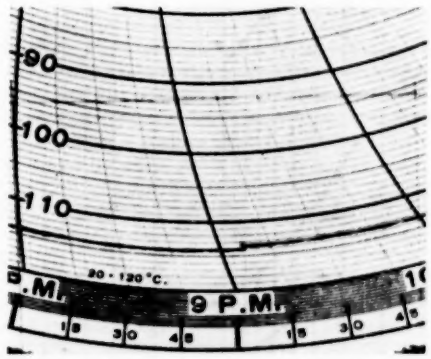


FIG. 9 EFFECT OF LOAD CHANGE ON EXIT TEMPERATURE OF HEAT EXCHANGER

Cascade-Control System. Up to this point, the effect of load disturbances only had been considered. It was known, of course, that the steam-supply pressure was liable to fluctuate widely; the effect of the maximum probable pressure fluctuation on the output temperature, when on control, was measured. The maximum deviation was 3 C. The merits of cascade control therefore were investigated; the obvious method was to arrange the output of the temperature controller to reset the desired value pointer of an instrument which measured and controlled the steam pressure in the jacket.

To investigate this possibility, the analyzer output was connected to the steam valve as before, and the analyzer input was derived from a steam-pressure transmitter. The data are shown in Fig. 10, curves A. In this range of periods, the contribution of valve motor and pressure-detecting element to the frequency response is negligible. The gain of the pressure controller was reduced from the limiting value, which caused continuous oscillations of the pressure control loop, to 20, so that the pressure would follow a sudden pointer movement without overshooting. Proportional control action was used.

The precise effect of introducing this cascaded loop into the temperature-control system was evaluated by drawing up a series of vector diagrams for a number of selected periods of oscillation. The example in Fig. 11 refers to a period of 30 sec. The vectors are θ_1—deviation; θ_2—pressure controller output (potential correction); θ_3—measured steam pressure; θ_4—temperature controller output (potential correction) applied to index pointer of pressure controller.

The ratio θ_2/θ_1 equals the gain of the pressure controller, and θ_3/θ_2 is the attenuation in that part of the plant involved in the pressure control loop. The vector diagram is constructed by drawing θ_1 of unit length and calculating the length of θ_3. The angle α is that which was measured in the steam-valve/steam-pressure analysis. θ_4 is the vector difference between deviation and measured variable. The difference between the angles α and β is the phase angle to be subtracted from the over-all phase lag (steam valve/liquor-temperature) previously determined; and the ratio θ_2/θ_3 is divided into the attenuation of the over-all system to find the frequency-response diagram relative to the temperature controller. The latter is shown in Fig. 10, curves B. This predicts a period of oscillation reduced from 54 to 39

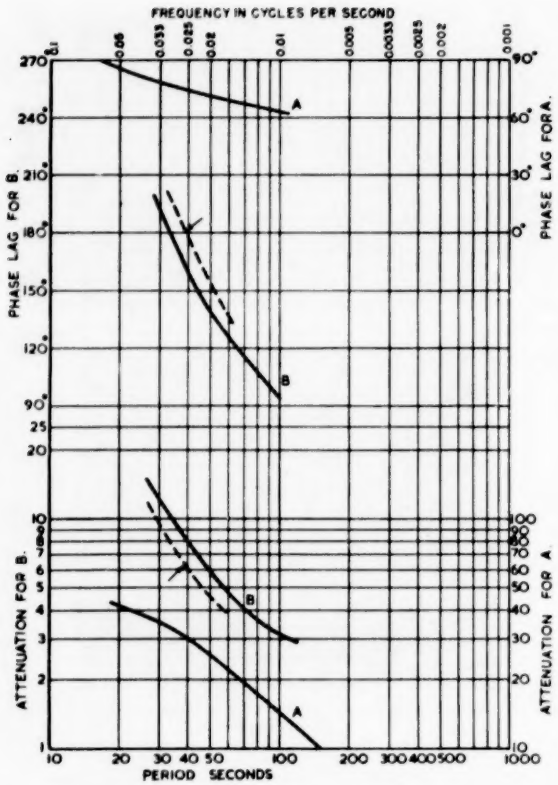


FIG. 10 FREQUENCY RESPONSE FOR CASCADE SYSTEM

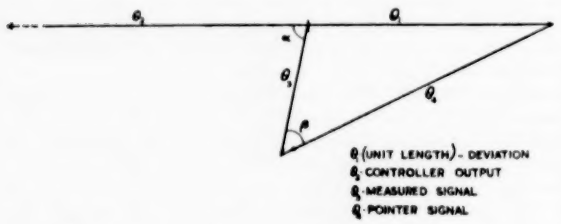


FIG. 11 VECTOR DIAGRAM FOR CASCADE SYSTEM

sec; the proportional band width required in the temperature controller depends on pressure-controller range, reset-mechanism range, the potential values of the plant, and the pressure-controller gain. It worked out at 4 C.

The practical value of the cascade control system was found to be considerable. Sustained disturbances in steam-supply pressure were reduced by the fast-acting pressure controller to short-duration impulses in steam-jacket pressure, before the liquor temperature was affected and the temperature controller came into action. The liquor temperature is now kept within the required limits at all times and, as a result of the shortened period, the effect of liquor-flow variations on the outlet temperature is less than it was when the simple control loop was used.

This method of assessing cascade control systems may appear to be rather ponderous, compared with the usual method of setting up the system and observing results, but it can be claimed to have very useful application in design work when the characteristics of the system are known from experience or calculation.

Controllability of Steam-Heated Heat Exchangers. It is pertinent to consider how far the practical data obtained in these experiments are in accord with theoretical calculations, and how useful these may be in predicting the performance of similar types of heat exchangers in the future.

A method is given in the Appendix for calculating the frequency response of the heat exchanger, which follows the work of Profos and Takahashi, and for which the author is indebted to a colleague (14). It is based on the simplifying assumptions that the tube-wall thermal capacity is negligible compared with the thermal capacity of the liquor in the tube, that the steam pressure is at all times equivalent to the saturated-steam temperature, and that the thermal capacity of the jacket wall is negligible compared with the effective thermal capacity of the steam and water.

It is further assumed that the conductances inside and outside the tube are the same for every tube.

The calculated frequency response was identical for temperatures at the 22nd tube and liquor exit. Individual measured values for the 22nd tube agreed with the calculated curve to within ± 10 deg phase and ± 30 per cent in attenuation over the limited range of periods considered (after subtracting the response of the detecting element from the measured values). The measured phase lag for exit temperature is greater than the calculated, by an amount which could be attributed to a distance-velocity lag of about 4 sec, but the measured attenuation agreed with calculated values within ± 30 per cent.

The cause of this extra phase lag is thought to be the initial superheat of the steam when it is let down through the control valve, and where it impinges at various points on the last tubes. The superheat amounts to 20 C under normal conditions, which would be sufficient to create unwetted segments on the last few tubes (to which heat transfer from steam would be very low) and thus cause the effect of a distance-velocity lag to appear in the control loop.

The heat exchanger was, in the usual sense, overdesigned, as the temperature difference at the liquor outlet was quite small. From the point of view of automatic control, however, this overdesign is not a disadvantage. It has been found in experiments on smaller-scale heat exchangers of this type (i.e., isothermal on the supply side) that for a constant supply-side capacity the deviation reduction factor for both demand-flow and inlet-temperature disturbances increases with increasing tube length. At one limit, as the tube length tends to zero, the deviation-reduction factor tends toward unity, for the effect of the disturbance passes out of the plant before its initial amplitude can be reduced. At the other limit, as the tubes become longer and the final temperature difference tends to zero, the deviation reduction factor for flow disturbances tends toward a limiting value, although the factor for temperature disturbances seems to increase indefinitely with tube length.

In this type of plant, the conventional block diagram with blocks labeled "supply side" and "demand side" is not a satisfactory representation, because the path followed by demand-fluid disturbances through the plant is not coincident with the path followed by potential correction, although closely related to it. In this plant a measured deviation reduction factor was about 60 for a demand-flow step disturbance; the calculated attenuation at the operating period of 54 sec is 44 (see Appendix). This method of estimation, although it cannot be claimed to be accurate, gives results of the right order of magnitude as a basis for exact calculations, but obviously there is a need for methods which are more exact and preferably no more complicated than the one used here.

Conclusion

The author has attempted to show, by describing two examples taken from practical experience of frequency-response methods, the value of these methods in the examination of process-control problems. Numerous authors of papers have advocated the systematic collection and publication of frequency-response data, and it is hoped that the exposition of the two plant investigations in this paper, however fragmentary and incomplete the data may be, will encourage further practical work and interpretation of data taken from actual plants. It can be claimed that the plant data obtained by the frequency-response method are very easy to manipulate, and that the practical technique is now firmly established and applicable without difficulty.

Three procedures are commonly used in the design of process-control systems to give adequate controllability. They are the investigation of the following:

(*a*) Alternative routes or paths for the potential correction to follow, in its progress through the plant.

(*b*) Or the application of individual control loops to every source of disturbance before these can enter the principal control loop.

(*c*) Or the adjustment of the plant design itself.

It is considered that procedure (*c*) has not been as fully exploited as it deserves.

The developments of improved methods for the calculation of controllability, and the correlation of these calculations with practical results, will in time enable process-control-system designers to predict the performance of these systems with an accuracy much higher than can now be attained. In a more positive sense, the system designer eventually will be able to influence the design of the plant itself so as to take the fullest advantage of the potentialities of automatic control.

BIBLIOGRAPHY

1 "Theoretical Foundations of the Automatic Regulation of Temperature," by A. Ivanoff, *Journal of the Institute of Fuel*, vol. 7, 1934, p. 117.

2 "Automatic Temperature Control of Jacketed Pans," by G. H. Farrington, Transactions of the Society of Instrument Technology, vol. 1, 1947, p. 2.

3 "The Practical Application of Frequency Response Analysis to Automatic Process Control," by C. I. Rutherford, Proceedings of The Institution of Mechanical Engineers, vol. 162, 1950, p. 334.

4 "A Portable Instrument for Process Control Analysis," by A. R. Aikman, *Instrument Practice*, vol. 5, 1951, p. 393.

5 "A Portable Instrument for Process Control Analysis. Part II: The Electrical Equipment," by J. R. Halsall, *Instrument Practice*, vol. 5, 1951, p. 793.

6 "Die Behandlung von Regelproblemen vermittels des Frequenzganges des Regelkreises," by P. Profos, Leeman, Zürich, Switzerland, 1943.

7 "Frequency-Response Analysis for Industrial Automatic-Control Systems," by D. W. St. Clair, W. F. Coombs, Jr., and W. D. Owens, Trans. ASME, vol. 74, 1952, pp. 1133–1150.

8 "Practical Application of Servomechanism Techniques to a Process Control Problem," by S. P. Higgins, Jr., and G. W. McKnight, *Instruments*, vol. 25, 1952, p. 777; also *Textile Research Journal*, vol 22, 1952, pp. 202–213.

9 "Transfer Function Analysis of Heat Exchange Processes," by Y. Takahashi, *Automatic and Manual Control*, Butterworths Publications, Ltd., London, England, 1952, p. 235.

10 "The Characteristics of Air-Operated Controllers," by A. R. Aikman and C. I. Rutherford, *Automatic and Manual Control*, Butterworths Publications, Ltd., London, England, 1952, p. 175.

11 "Analysis of Pneumatic Controllers," by J. M. L. Janssen, *Automatic and Manual Control*, Butterworths Publications, Ltd., London, 1952, p. 189.

12 "The Development of Automatic Process Control," by A. J. Young, Transactions of Instruments and Measurements Conference, Stockholm, 1952 (to be published).

13 "The Calculation of Process Control Settings From Frequency Characteristics," by N. Ream (to be published in the Transactions of the Society of Instrument Technology).

14 Private Communication from I. Gray (I.C.I. Ltd.).

Appendix

Terminology

The following terms used in the text are not defined in British Standard 1523:1949. This Standard is now under revision, and these terms will possibly be included in the revised edition. They are presented in the order in which they occur in the text.

Amplitude Ratio. The amplitude of a signal that must be applied to the regulating unit to give rise to a measured oscillation of unit amplitude at any particular period. This may be expressed in any convenient units (e.g., psi/deg C).

Attenuation. The amplitude of potential value that must be applied to the plant to give rise to a measured oscillation of unit amplitude at any particular period. (This is dimensionless.)

Subsidence Ratio. The ratio of the amplitudes of the first two consecutive peaks of a damped oscillation (measured in the same direction) from the mean about which the oscillation is taking place.

Operating Period. The period of any oscillations of the controlled condition at which the system operates, following a disturbance.

Controller Gain. The amplitude of the potential correction that the controller applies to the process per unit amplitude of deviation measured at any particular period (dimensionless).

Nominal Action Times. The values of derivative action time and integral action time corresponding to the time constants of the instrument, and which manifest themselves when either integral or derivative action (but not both) is used with proportional action.

Proportional Control Factor. Present definition 2505 may be amended to read, "That part of the controller gain due to proportional control action only."

Interaction Factors. The factors which relate any parameter in the transfer function of a controller to the nominal values of action times and proportional band.

Disturbance. Any condition or alteration of condition imposed on the system which tends to affect the value of the controlled condition.

Deviation Reduction Factor. The factor by which the amplitude of measured deviation is less than the amplitude of potential deviation applied, due to the operation of the controller. (This factor depends on the form and point of occurrence of the disturbance which gives rise to the potential deviation.)

Correction of Frequency Response for Subsiding Oscillations

A method of correcting the frequency-response characteristic of a plant, in order to ascertain the controller settings for subsiding oscillations, has been derived from fundamental considerations (13). The method can be used for any desired subsidence ratio. For very many cases of automatic process control a subsidence ratio of $\epsilon : 1$ is satisfactory. A simple practical method is to measure the slope of the attenuation and phase curves with a protractor and then make the corrections according to Table 1.

The correction to the phase curve is always positive, and the correction to the attenuation curve always negative.

The values given in this table apply to the standard graph paper which is in use in I.C.I., Ltd., and has the following scales:

Log period: 1 cycle = 3 in.
Log attenuation: 1 cycle = 3 in
Phase lag: 40 deg = 1 in.

TABLE 1 PHASE AND ATTENUATION CORRECTIONS

Slope of attenuation curve, deg	−10	−20	−30	−40	−50	−60	−70	−80
Phase correction, deg	2	3	5	8	11	16	25	52
Slope of phase curve, deg	−10	−20	−30	−40	−50	−60	−70	−80
Attenuation correction, in.	− 0.03	− 0.07	− 0.11	− 0.16	− 0.20	− 0.33	− 0.52	− 1.06

Theoretical Calculation of Frequency Response of Heat Exchanger

The theoretical analysis has been based on the following equations

$$\frac{\partial\theta}{\partial t} + v\frac{\partial\theta}{\partial x} = \frac{\alpha U}{w}(\phi - \theta) \quad \ldots\ldots [1]$$

and

$$C\frac{d\phi}{dt} = qL + \int_0^l \alpha U(\theta - \phi)dx \quad \ldots\ldots [2]$$

where

$\theta(x, t)$ = fluid temperature, deg C
$\phi(t)$ = steam temperature in jacket, deg C
v = fluid velocity, cm/sec
$q(t)$ = steam flow, kg/cm
C = equivalent heat capacity of steam space, cal/deg C
L = latent heat of steam, cal/kg
α = over-all heat-transfer coefficient, cal/cm² deg C sec
l = total tube length, cm
U = tube circumference, cm
w = heat capacity of fluid flow per unit length, cal/deg C cm
x = distance along tube, cm
t = time, sec

Equation [1] is the partial differential equation, given by Profos (6), expressing the heat exchange between a flow and a solid surface, and Equation [2] is a natural development from this. The derivation of these equations is based on considerations of heat balance, and the reader is referred to Profos for further details. However, the following basic assumptions which have been made should be noted:

1 Heat flow in the region consisting of the tube and fluid content takes place in the radial direction only, except for transport; i.e., there is no heat flow by conduction in the direction of the axis of the tube, either within the fluid or in the tube wall itself.

2 The fluid temperature θ is isothermal over any cross section of the flow.

3 The steam temperature ϕ is uniform throughout the jacket.

4 The rate of increase of steam temperature, ignoring any heat losses, is equal to the latent-heat input divided by the equivalent heat capacity of the jacket contents. Sensible heat is neglected.

5 All solid heat capacities are neglected.

6 System parameters are constant, including the heat-transfer coefficient α, and in the first instance, the fluid velocity v.

7 There are no heat losses.

In view of the nature of the assumptions made, particularly 3, 4, and 6, the analysis must at the best be considered as only giving a rough guide to the behavior of the system. It is possible to take into account the effects of the solid heat capacities (in the tube wall, and so on), but the refinement is probably not justified in this particular case; the uncertainties in the assumptions made concerning the distribution of steam temperature and its dependence on steam flow will still cast considerable doubt on the accuracy of the results. However, until such time as these matters have been investigated fully, the following type of analysis should indicate the nature and order of magnitude of the effects to be observed:

Response to Oscillations in Steam Flow. Equations [1] and [2] may be rewritten in more convenient form in the following manner:

Put

$$\left.\begin{aligned} T &= tv/l \\ y &= x/l \\ a_1 &= \alpha Ul/wv \\ a_2 &= \alpha Ul^2/Cv \end{aligned}\right\} \text{ dimensionless}$$

$$\psi = q\frac{L}{Cv_1} \text{ deg C}$$

The equations become

$$\frac{\partial\theta}{\partial T} + \frac{\partial\theta}{\partial y} = a_1(\phi - \theta) \quad \ldots\ldots [3]$$

and

$$\frac{d\phi}{dT} = \psi + \int_0^1 a_2(\theta - \phi)dy \quad \ldots\ldots [4]$$

From these two equations, ϕ may be eliminated, and the following response in θ corresponding to a sinusoidal oscillation in steam flow may be derived:

$$\theta(y, T) = \frac{a_1(a_1 + j\Omega)(1 - e^{-a_1 y}e^{-j\Omega y})e^{j\Omega T}}{\Phi(j\Omega)} \text{ (real part)} \quad \ldots [5]$$

where

Ω = dimensionless frequency = $\omega l/v$
ω = frequency of oscillation of steam flow
$j = \sqrt{-1}$

and

$$\Phi(j\Omega) = (a_1 + j\Omega)^2(a_2 + j\Omega) - a_1a_2(a_1 + j\Omega) + a_1a_2(1 - e^{-a_1}e^{-j\Omega}) \quad \ldots\ldots [6]$$

To calculate the response in θ at any point along the tube, numerical values for a_1, a_2, and y are inserted, and the attenuation and phase lag calculated for a range of frequency Ω, by determining the modulus and argument of the frequency Function [5].

Taking the following values:

v = 17.2 cm/sec
C = 2 × 10⁴ cal/deg C
α = 0.0246 cal/cm deg C sec
l = 5500 cm
U = 12.3 cm
w = 12.1 cal/cm deg C
$x = l$ = 5500 cm.

that is

a_1 = 8.0
a_2 = 26.6
y = 1

the following values for attenuation and phase lag were calculated:

Period of oscillation, sec	125	62.8	31.4	12.5
Ω	16	32	64	160
Attenuation A	22.5	57.4	175	394
Phase lag, deg	115	133	151	168

(The attenuation given here is the dimensionless attenuation which becomes unity at infinite period, $\Omega = 0$.)

Response to Oscillations in Fluid Velocity. For changes in fluid velocity v, Equation [1] becomes nonlinear. However, a simple approximate solution can be obtained if only small changes in v are considered. Put

$$\left.\begin{aligned} v(t) &= v_0 + v_1(t) \\ \theta(x, t) &= \theta(x) - \theta_1(x, t) \\ \phi(t) &= \phi_0 - \phi(t) \\ q &= q_0 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [7]$$

where v_0 is the steady velocity about which small variations v_1 are taking place, and θ_1, ϕ_1 are corresponding small fluctuations in temperature about means θ_0 and ϕ_0. The steam flow is supposed to be held constant at the steady value q_0.

Substituting the expressions given by Equations [7] into Equations [1] and [2] gives

$$-\frac{\partial \theta_1}{\partial t} + (v_0 + v_1)\left(\frac{d\theta_0}{dx} - \frac{\partial \theta_1}{\partial x}\right) = \frac{\alpha U}{w}(\phi_0 - \phi_1 - \theta_0 + \theta_1) \quad \ldots\ldots [8]$$

and

$$-C\frac{d\phi_1}{dt} = q_0 L + \int_0^l \alpha U(\theta_0 - \theta_1 - \phi_0 + \phi_1)dx \quad \ldots [9]$$

Equating quantities of like order in Equations [8] and [9], and ignoring the second-order product $v_1(\partial\theta_1/\partial x)$ gives rise to the equations

$$\frac{\partial \theta_1}{\partial t} + v_0\frac{\partial \theta_1}{\partial x} = \frac{\alpha U}{w}(\phi_1 - \theta_1) + v_1\frac{d\theta_0}{dx} \quad \ldots\ldots [10]$$

$$C\frac{d\phi_1}{dt} = \int_0^l \alpha U(\phi_1 - \theta_1)dx \quad \ldots\ldots\ldots\ldots [11]$$

$$v_0\frac{d\theta_0}{dx} = \frac{\alpha U}{w}(\phi_0 - \theta_0) \quad \ldots\ldots\ldots\ldots [12]$$

and

$$0 = q_0 L + \int_0^l \alpha U(\theta_0 - \phi_0)dx \quad \ldots\ldots\ldots\ldots [13]$$

Equations [12] and [13] represent the steady-state heat balance, from which an expression for $(d\theta_0)/(dx)$ may be derived for substitution in Equation [10]. Again, the equations may be made nondimensional

$$\frac{\partial \theta_1}{\partial T} + \frac{\partial \theta_1}{\partial y} = a_1(\phi_1 - \theta_1) + r\frac{d\theta_0}{dy} \quad \ldots\ldots [14]$$

$$\frac{d\phi_1}{dT} = a_2\int_0^1 (\theta_1 - \phi_1)dy \quad \ldots\ldots\ldots\ldots [15]$$

$$\frac{d\theta_0}{dy} = a_1(\phi_0 - \theta_0) \quad \ldots\ldots\ldots\ldots [16]$$

and

$$0 = \psi_0 + a_2\int_0^1 (\theta_0 - \phi_0)dy \quad \ldots\ldots\ldots\ldots [17]$$

where

$$\left.\begin{aligned} y &= x/l \\ T &= tv_0/l \\ a_1 &= \alpha Ul/wv_0 \\ a_2 &= \alpha Ul^2/Cv_0 \\ r &= v_1/v_0 \end{aligned}\right\} \text{dimensionless}$$

and

$$\psi_0 = q_0 lL/Cv_0 \text{ deg C}$$

and by eliminating ϕ_1 from Equations [14] and [15], the response in θ_1 to oscillations in v_1 (that is, r) may be derived

$$\theta_1(x, T) = \frac{\psi_0 a_1^3 e^{-a_1 y}}{a_2(1 - e^{-a_1})}\frac{(a_1 + j\Omega)(a_2 + j\Omega)(1 - e^{-a_1 y}e^{-j\Omega y})}{\Phi(j\Omega)}$$

where $\Phi(j\Omega)$ and Ω are as in (A).

Taking the same values for the parameters as before, and with $v_0 = 17.2$ cm/sec, the following values of attenuation and phase lag were calculated:

Period of oscillation, sec	125	62.8	31.4	12.5
Ω	16	32	64	160
Attenuation A	19.2	36.7	67	64.6
Phase lag, deg	56	93	128	159

(Again A is the dimensionless attenuation which becomes unity at infinite period.)

Discussion

P. S. Buckley.[3] This paper represents a valuable addition to the growing number of process-control studies which have been appearing in the technical literature in the past few years. Like many of its recent predecessors, it is based on what is known as the "frequency-response" method of analysis. The point of view taken by the author is not that of a theorist, but rather that of a practical process-control engineer who is confronted with problems in an existing chemical plant.

Of the two examples chosen by the author, one, the automatic control of a heat exchanger, is of very broad interest as it is a frequently encountered problem in chemical and petroleum plants. The approach taken is partly empirical, partly theoretical. The author determined the transfer functions or frequency responses of the process and instrumentation experimentally, checked the results partially by theory, and calculated necessary controller settings to give desired controllability.

In the course of reviewing a number of European papers on process control, this writer has been struck by the differences in European and American practice in such matters as terminology and graphical representation. This is at times confusing and therefore it is suggested that there might be merit in attempting to work out standardization.

J. M. L. Janssen.[4] In his introduction the author states that one of the main purposes of dynamic plant analysis is the accurate prediction of control-system performance at the design stage. I wonder how far that can go. As the author has emphasized, information on magnitude, form, and location of disturbances is at least as important as that on the over-all characteristics of the plant. Although the step disturbance is very popular for characterizing control-system performance, actual disturbances may be somewhat different. Fig. 3, for instance, gives a pattern of deviations that does not appear to be caused by isolated step disturbances. The accurate prediction would require a priori knowledge of the actual disturbances, and without that I

[3] Engineer, E. I. du Pont de Nemours & Co., Orange, Texas.

[4] Research Engineer, Royal-Dutch/Shell-Laboratory, Delft, Holland.

must confess I am rather pessimistic as to the possibility of a really accurate prediction. Perhaps the difficulty lies in the terminology, and I should ask what is meant here by the term "accurate."

Elsewhere in the paper the author deals with the difficulty of the path followed by the potential correction being different from the one followed by the demand disturbances. Out of many alternatives the block diagram in Fig. 12 of this discussion may be suggested for those cases.

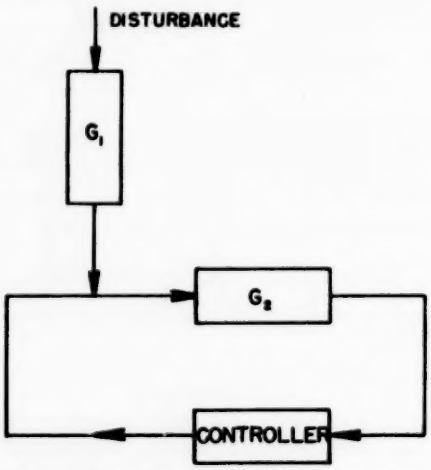


FIG. 12 BLOCK DIAGRAM FOR HEAT EXCHANGER

For the example of the heat exchanger, G_2 represents the transfer function of the heat exchanger for potential corrections. The product G_1G_2 represents the transfer function for demand disturbances. Both can be measured or perhaps be calculated. If G_2 and the product G_1G_2 are known, then G_1 is easily found. Block G_1 has no physical reality but for the purpose of analysis that does not matter.

G_1 modifies the shape of the disturbances and when G_1 is approximately equal to unity over a sufficiently wide frequency range, then block G_1 can be omitted and the deviation reduction factor can be calculated along the regular lines. But if, for instance, the response of the heat exchanger to demand changes is much slower than to changes in valve position, then G_1 is a block with time lags and the deviation reduction factor must be expected to have a higher value. This was probably the case in the author's example. In so far as I can see, it does not really matter whether or not the system has distributed parameters. A similar difference in path can exist in a system with lumped parameters.

E. G. HOLZMANN.[b] Control engineers are well aware that process-control loops, in general, are a class of regulating systems rather than of servomechanisms. The author illustrates the distinction by stressing the fact that the proper measure for quality of control in a chemical plant is the ability of the control system to reduce the effects of disturbances. He rightly emphasizes the influence of the location of the disturbances in the loop.

One would like to add that, where possible, the time pattern of the disturbances should be taken into account. Rapid, but small, fluctuations of the controlled quality often can be tolerated although they may spoil the appearance of a control record. Sometimes good plant operation requires that the controlled response in one control loop be slowed down deliberately in order to minimize undesirable effects on some other part of the process. These effects may be due to variations in mass or energy flow dictated by the sensitive control action of a control loop adjusted for maximum speed of response. Where control is effected through the set point of a controller, as, for example, in a cascade system, high speed of response is undoubtedly desirable. But in the majority of applications the set point is held constant. It would seem that in many such cases tuning to a subsidence ratio of ϵ is unnecessary and possibly undesirable.

The author's mastery of the techniques of analysis employed in his paper tends to obscure the over-all pattern of the analysis. Four major steps are involved, viz.:

1 Determination of frequency response of plant.
2 Modification for chosen subsidence ratio.
3 Calculation of controller phase angle for chosen ratios of integral and derivative action times to operating period.
4 Determination of operating period and of the corresponding plant attenuation.

The analysis insures that the system is stable, with a certain amount of damping, and gives (*a*) the required proportional band, and (*b*) the period of damped oscillation of the control loop.

Comparing the procedure with the more familiar approach based on the phase-margin criterion, one sees that essentially the same results may be obtained, but that the second step, that of modifying the frequency-response curves, is saved. Perhaps more significant than the saving in labor is the perspicuity of the phase-margin approach which permits one not only to obtain the answer for a single predetermined mode of operation, but also to get an insight into the behavior under nonideal conditions of controller adjustment.

AUTHOR'S CLOSURE

The author, like Mr. Buckley, has been struck by the differences between European and American terminology. It is certainly hoped that a larger area of agreement between rival terminologies can be reached in the near future, and that the present anomalous position, in which there are different terms in the same language for the same basic concepts, will be rectified.

The graphical representation used in this paper is indicative only of the practice in the author's company and is not to be taken as representative of the European practice as a whole; if terminology could be unified the interpretation of graphs would present little difficulty.

It is the case, as Mr. Janssen points out, that prediction of system performance requires a priori knowledge of the pattern of the actual disturbances, as well as their points of entry. The accuracy of prediction will be no greater than the accuracy with which disturbance patterns are known. In existing plant, it is sometimes possible to measure disturbances at their points of entry: this was the case in the second example chosen by the author, but not in the first. If a system is at the design stage, disturbance patterns must be estimated either from knowledge of the operation of the plant or from previous experience on similar plants; the assumption of step disturbances in every case can lead to very conservative design. There is no doubt that much remains to be done in the systematic collection and analysis of disturbance patterns and their correlation with system performance.

Block-diagram systems, such as Mr. Janssen suggests, are helpful abstractions provided that the use of them does not induce the unwary to ascribe the blocks to particular sections of the physical plant.

Adjustment of the controller to give a subsidence ratio of ϵ:1 (corresponding to a maximum value of deviation ratio of about

[b] Engineer, Shell Development Company, Emeryville, Calif.

3) was not suggested by the author as being the optimum condition in every application; it is better regarded as a limiting condition which may have to be modified in some cases by such practical considerations as Mr. Holzmann suggests.

The approach to controller adjustment based on phase-margin, and similar techniques, does allow a deeper insight into control-system behavior than the complementary approach used by the author in this paper, but if the objective is limited to the adjustment of a controller (of previously established and tabulated characteristics) to obtain a definite subsidence ratio, then the method used by the author enables this objective to be attained with much less labor than it may appear.

The author is indebted to Dr. R. Oldenburger for the invitation to contribute to the Symposium, and to his head of department, Mr. A. J. Young, without whose leadership this work would not have been possible.

Frequency-Response Methods Applied to the Study of Turbine Regulation in the Swedish Power System

By V. OJA,[1] NEW YORK, N. Y.

This paper describes principles worked out for the design of frequency-power regulation in a power system. The method of analysis is based upon general ideas of frequency-response methods and offers new possibilities for evaluating the regulation properties of power stations. It is aimed to give efficient regulation with the least possible effort. This is attained by matching the feedback data for the governors to the characteristics of the machines and the actual disturbances in the network. This method has been worked out for the Swedish power system. The power system is described briefly, the characteristics of the elements in the regulation loop are given, and the measurements of the frequency-response curve of the power system are reviewed.

Nomenclature

The following nomenclature is used in the paper:

Time Variables. The following quantities refer to sinusoidal variations superimposed on the mean value. They are vector quantities but since they are not used with any other significance in this paper the vector signs are omitted:

a' = position of auxiliary servomotor
a = relative gate opening
f = frequency of power system
P = power
P_T = turbine output power
P_s = power disturbance
P_2 = residual disturbance (after regulation)

Transfer Functions. These are complex, frequency-dependent functions relating the vector expressions for the time variables:

F = transfer function for element in regulation system
$|F|$ = absolute value of F
φ = phase angle of F
Φ_k = transfer function for open-loop with damping
Φ_0 = transfer function for open-loop without damping

Other Quantities:

α_b = mean gate opening
τ_0 = time constant of regulator
τ_2 = starting time of water
τ_c = time constant of stabilizing feedback
δ_p = permanent speed droop
δ_t = transient speed droop
η_r = regulation efficiency
β_r = relative regulating effort
ω = angular frequency of disturbance
ρ = rigidity of network, mw/cycles/sec

Swedish Network

The whole of Sweden is covered by a jointly operated network, its geographical extent being some 900 miles North to South and 300 miles East to West. The generating capacity of the system is some 3500 mw (1949). Water power predominates producing 97 to 98 per cent of the total energy during normal years. About 85 per cent of the hydroelectric power is generated in the northern half of the country, where there is only one sixth of the population, while the greatest consumer centers are located in the South. The power is transmitted by a system of 380 and 220-kv lines forming the backbone of the network, as shown in Fig. 1.

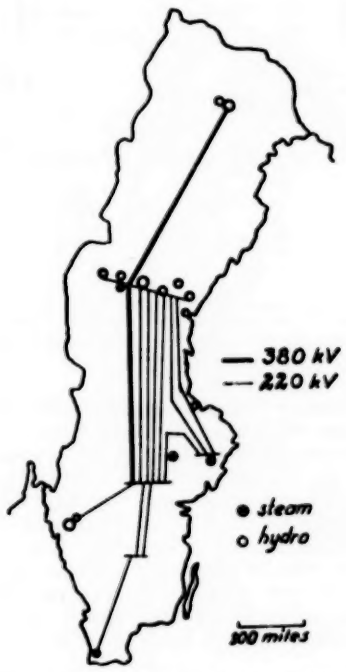


Fig. 1 Schematic View of Swedish 380 and 220-Kv Network With Principal Power Stations

The consumption can be divided roughly into 65 per cent industry, 25 per cent domestic, and 10 per cent railroad load.

The planning of frequency-power regulation is effected in Sweden by the State Power Board. With the given configuration

[1] Former member of Research Department, Swedish Electric Company, ASEA, Sweden; now with its United States subsidiary, Aros Electric, Inc.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 11, 1953. Paper No. 53—A-23.

of the network, the following scheme is under development (1):[2]

One large plant, the master station, with an adequate water reserve, is assigned to take care of big and slow load variations. It operates astatically (flat frequency control) with very heavy damping. The response thus being slowed down, it is possible to operate the correct number of units in this station, maintaining good over-all efficiency and limiting wear.

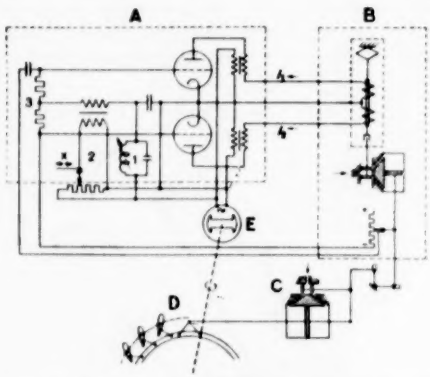
Numerous supporting stations are assigned to operate with speed droop and low damping. The speed droop in these stations should be combined against the output power, the water flow or its mean value, water level, steam pressure in thermal power plants or other quantities according to local conditions. These stations should take care of rapid load variations, working within a narrow band around their designated operating point.

As this scheme calls for great flexibility of the governors in order to combine different quantities in different proportions and permit a flexible damping adjustment, a brief description of the more recently designed governor will be given before a general analysis is undertaken.

Governor Operation

In all recently erected stations of the network, electrically operated turbine governors are used (2, 6).

The mode of operation is described briefly with reference to Fig. 2. The electron-tube regulator A is designed to analyze,



A, Electronic valve regulator; B, control mechanism; C, main servomotor; D, gate mechanism; E, auxiliary a-c generator.

Fig. 2 Schematic Diagram of Electrohydraulic Turbine Governor

combine, and to some extent amplify the control quantities and to convert them to direct current. The frequency sensitivity is obtained by the use of a tuned LC circuit A1. Several circuits A2 are used for the different quantities X employed for regulation. An RC circuit A3 provides adequate damping. The proportions of these different quantities can be adjusted readily by means of simple electrical rheostats at the control desk. The output current of the regulator actuates the coil in the electrohydraulic control unit B, which converts the signals from electric current to mechanical movement. The functions of the main servo C, the gate mechanism D, and the auxiliary generator E are self-explanatory.

Closed-Loop System

Assuming that only one machine in the network is regulating, the actuation of the governor being by the frequency and the gate opening only, the block diagram in Fig. 3 can be computed. A

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

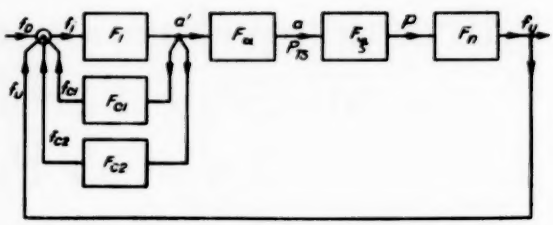


Fig. 3 Block Diagram of the Regulating System

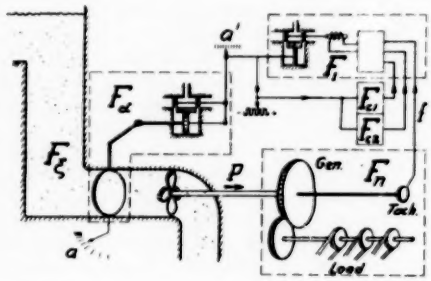


Fig. 4 Schematic Pictorial View of Regulating System

schematic view of the regulating system is shown in Fig. 4.

The first link in Fig. 3 describes the electrohydraulic turbine governor. F_1 being the transfer function of the regulator, combining frequency deviation f with the mechanical output a', can be expressed by

$$F_1 = \frac{a'}{f_i} = \frac{1}{j\omega\tau_0} \quad \ldots\ldots [1]$$

In the same manner, the permanent droop will be expressed by

$$F_{c1} = \frac{f_{c1}}{a'} = \delta_p \quad \ldots\ldots [2]$$

The output of this link, for reasons of convenience, has been expressed in terms of frequency deviation.

The transient speed droop or damping circuit is characterized by

$$F_{c2} = \frac{f_{c2}}{a'} = \delta_t \tau_c \frac{j\omega}{1 + j\omega\tau_c} \quad \ldots\ldots [3]$$

τ_c being the time constant of the RC circuit.

The link F_a represents the properties of the main servomotors as well as the turbine itself under normal water pressure (3). Thus a represents an equivalent gate opening which is proportional to the turbine power P_{T_3} at normal water pressure. This function in some cases can be expressed by analytical relationship or can be evaluated point by point.

The influence of the water inertia F_ξ for moderate heads and small deviations from a mean value, can be expressed by

$$F_\xi = \frac{P}{P_{T_3}} = \frac{1 - j\omega\,\tau_2\,\alpha_b}{1 + j\omega\,0.5\,\tau_2\,\alpha_b} \quad \ldots\ldots [4]$$

where τ_2 = starting time of water, and α_b = mean value of gate opening.

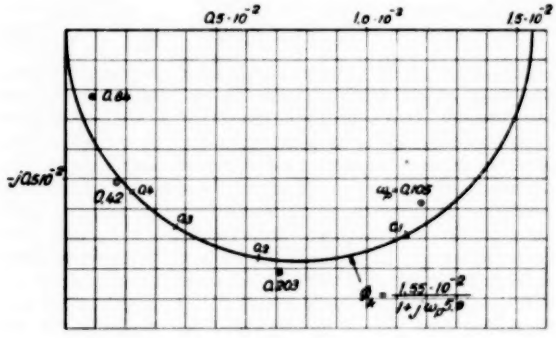
The quantity P thus is the real power output of the machine. The last link F_n represents the power system.

Transfer Function of the Network

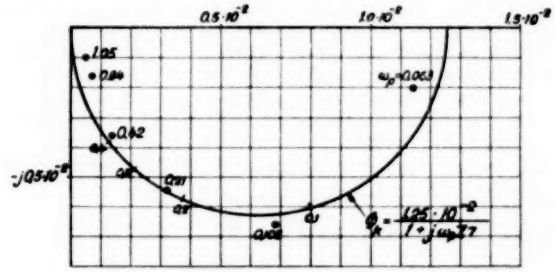
This function (4) is defined as the relationship between the

variations of the supply frequency f when the input power P to the system is varied. In order to obtain information about this relationship the transfer function actually was measured. During these measurements, the governors of all the turbines in the network were locked firmly and the frequency was held by manual regulation. The output of one 50-mw machine was then varied sinusoidally at different frequencies. The resulting frequency variations were measured and compared as to amplitude and phase to the varied input power in a specially designed measuring circuit. In order to separate the forced variations from the much greater random variations of the frequency, resulting from normal load disturbances, the measured quantities were multiplied by a sinus function with a frequency of the forced power variations. The results of this multiplication were integrated over a sufficiently long time (up to 10 min), thus eliminating the effect of the random variations.

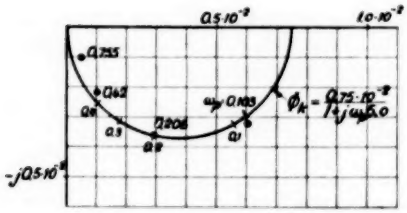
The results are shown as Nyquist plots in Figs. 5(*a*, *b*, *c*) for different load conditions in the network at different times of the day. In the same figures a curve is drawn corresponding to equation



(*a*)



(*b*)



(*c*)

FIG. 5 NYQUIST CURVES OF NETWORK IN CYCLES PER SEC/MW (*a*, Sunday, noon to 3 pm; *b*, weekday, midnight to 4 am; *c*, weekday, 10 am to 11 am.)

$$F_n = \frac{1}{\rho}\,\frac{1}{1 + j\omega\tau_n} \qquad [5]$$

Equation [5] gives fairly good values at lower frequencies. The discrepancy at higher frequencies is mainly because of the reactances in the transformers and the high-voltage lines at the point of measurement. However, as the gain is low at these higher frequencies, the simple expression can be used for preliminary analysis at all significant parts of the network.

METHOD OF ANALYSIS

Having thus defined all the elements in the closed-loop system, a method of analysis now can be worked out giving the data needed to make the units in the different stations work according to the prescribed scheme.

Following the general idea of frequency-response analysis, the disturbance is assumed to be a sine wave. Knowing the transfer functions of the elements, the response of the whole system can be computed readily. On the other hand, as the regulation is connected with wear and sometimes with efficiency drop, the analysis should serve to indicate a way of using given machines for regulation in an advantageous and economical manner.

On the basis of these concepts, the regulation will be characterized by two quantities, the efficiency of the regulation and the associated regulating effort.

The *efficiency* is defined as the ratio of the reduction in the amplitude of the disturbance to the original amplitude of the disturbance, i.e., as the actual improvement obtained expressed as a ratio. Fig. 6 shows a turbine with its governor and servos having an over-all transfer function F_T, connected to a network F_n.

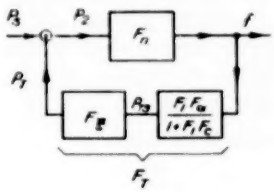


FIG. 6 BLOCK DIAGRAM WITH ONE REGULATING MACHINE

The obvious relation

$$P_2 = P_s - P_T \qquad [6]$$

is obtained, and the actual improvement is

$$|P_\eta| = |P_s| - |P_2| \qquad [7]$$

According to definition, the regulating efficiency is

$$\eta_r = \frac{|P_\eta|}{|P_s|} \qquad [8]$$

By insertion of the transfer function of the open loop

$$\Phi_k = \frac{P_T}{P_2}$$

$$\eta_r = 1 - \left|\frac{1}{1 + \Phi_k}\right| \qquad [8a]$$

is obtained.

It is thus possible to derive a curve $\eta_r(\omega)$ which shows the efficiency of the arrangement with a sinusoidal disturbance of different frequencies ω. The regulation efficiency becomes

$\eta_r = 1$ if entire disturbance is eliminated
$\eta_r = 0$ if amplitude of disturbance is unchanged
$\eta_r < 0$ if regulation increases disturbance

The two latter cases obtain if the turbine power P_T is out of phase in relation to the disturbance.

The second characteristic is the *regulating effort* which must be expended to obtain the regulation efficiency derived in the foregoing. The regulating effort is defined as the ratio of the absolute values of the change in gate opening and the amplitude of the disturbance. In Fig. 6 the change in gate opening is expressed as the static power P_{Ts}.

According to the definition, the regulating effort is obtained as

$$\beta_r = \frac{|P_{Ts}|}{|P_s|} \quad [9]$$

If substitution is made of the transfer function for the water pressure and the open-loop transfer function is introduced, the following equation is obtained

$$\beta_r = \left| \frac{1}{F_\xi} \frac{\Phi_k}{1 + \Phi_k} \right| \quad [10]$$

It is thus possible to derive a curve β_r (ω) which gives the amplitude of the gate movement under a sinusoidal disturbance of unit amplitude.

An example will help to show the application of these quantities. The machine used in the calculation is a 50-mw Kaplan turbine with a head of about 100 ft which has been investigated experimentally. The water-starting time is $\tau_2 = 1.5$ sec, and the load on the network corresponds to that of the Sunday curve.

Evaluation of Feedback Constants

We will assume that only one machine is regulating and, referring to Fig. 3, that the connection F_{e1} is opened and astatic regulation is obtained. The transfer function of the main branch without feedback F_{e2} then will be

$$\Phi_0 = F_1 F_a F_\xi F_n \quad [11]$$

becoming with the damping feedback F_{e2} connected

$$\Phi_k = \frac{\Phi_0}{1 + F_1 F_{e2}} \quad [12]$$

where, with reference to Equations [1] and [3]

$$F_1 F_{e2} = \frac{1}{\tau_0} \frac{\delta_t \tau_e}{1 + j\omega\tau_e} \quad [13]$$

In these equations, all the quantities are known except the damping constants δ_t and τ_e. We will assume values for these constants taken from an actual test, giving fairly good but "lively" regulation. We take $\delta_t = 1$ per cent and $\tau_e = 6.5$ sec.

The frequency-response curve of the open loop Φ_k then is defined. With this curve as a basis, curves for efficiency and regulating effort are derived using Equations [8a] and [10]. Results are shown as curves 1 in Fig. 7.

It will be noted that the regulation efficiency becomes negative for $\omega > 0.17$, i.e., for a disturbance-frequency of about 1.7 cycles per min and reaches a value of approximately —1 for $\omega = 0.3$ while the regulation effort is still high.

According to the scheme described, the astatically regulated master station should be damped quite heavily. A method will be described which gives the damping constants starting from some predetermined conditions for the efficiency curve η_r (ω). As an example, the simple condition of fixing the minimum value of the efficiency to —0.2 will be treated. This means according to Equation [8a]

$$-0.2 \leq 1 - \left| \frac{1}{1 + \Phi_k} \right| \quad [14]$$

and thus

$$|1 + \Phi_k| \geq 0.83 \quad [14a]$$

Further calculation is made graphically and is shown in Fig. 8. The Φ_k curve with feedback, according to Equation [14a], must pass the point —1 at a distance not less than 0.83 and thus must lie outside a circle having as its center the point —1 and radius 0.83; the Φ_k curve being obtained from the Equation [12].

At any point on the known Φ_0-curve, for example, $\omega_1 = 0.2$, it is possible to define certain conditions which the function $1 + F_1 F_{e2}$ must fulfill at this frequency. If the vector $1 + F_1 F_{e2}$ at $\omega = 0.2$ has a phase angle of 0 deg then the vector Φ_k will have the same angle as Φ_0 (in this case 168 deg) and an amplitude which is decided by the absolute value of $1 + F_1 F_{e2}$. In order that the corresponding point on the Φ_k-curve shall not appear within the circle drawn, the absolute value of $1 + F_1 F_{e2}$ may not lie between 1.34 and 13.4; these limits are obtained from the ratios $\overline{AO}/\overline{BO}$ and $\overline{AO}/\overline{CO}$. If, instead, the vector $1 + F_1 F_{e2}$ at $\omega = 0.2$ has an angle of, say, 40 deg, the amplitude may not lie within an area limited in one direction by 6.2 ($\overline{AO}/\overline{EO}$). In this way the points B', C', and E' are obtained through which a circle may be constructed to encompass those values which the function $1 + F_1 F_{e2}$ may not assume at a frequency $\omega = 0.2$ if it is to fulfill the original requirements as specified.

It can be shown easily that the diameter of the circle thus obtained is

$$D = \overline{G'F'} = \frac{\overline{AO}}{\overline{GO}} - \frac{\overline{AO}}{\overline{FO}} = |\Phi_0| \left[\frac{1}{1 - \gamma} - \frac{1}{1 + \gamma} \right] = |\Phi_0| \frac{2\gamma}{1 - \gamma^2} \quad [15]$$

where γ = radius of the original circle (0.83 in this case).

The center of this circle lies on the prolongation of OA at a distance from the origin equal to

$$\overline{OO'} = \frac{\overline{AO}}{\overline{FO}} + \frac{D}{2} = |\Phi_0| \left[\frac{1}{1 + \gamma} + \frac{\gamma}{1 - \gamma^2} \right] = \frac{|\Phi_0|}{1 - \gamma^2} \quad [16]$$

This example construction for $\omega = 0.2$ is repeated for a number of frequencies to give a host of circles with different frequency designations. The feedback function $1 + F_1 F_{e2}$ should be chosen so that the point for a frequency ω_s shall lie outside its corresponding circle. In the present case the feedback function $1 + F_1 F_{e2}$ is represented according to Equation [13], by a semicircle

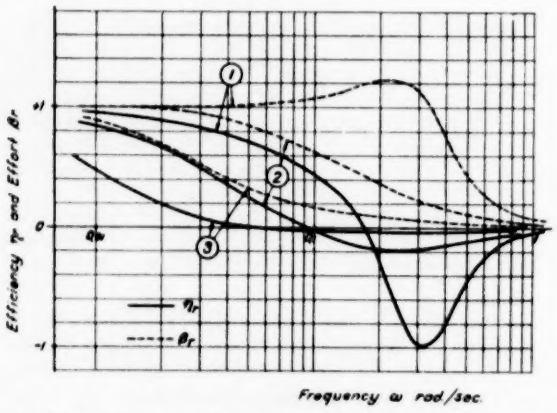


Fig. 7 Efficiency and Regulating Effort for One Machine With No Permanent Speed Droop

with its center on the real axis and, for $\omega = \infty$, by the point +1. The radius of the semicircle is chosen in such a way that it is the smallest possible which will comply with the requirements mentioned. This semicircle in Fig. 8 should pass in the immediate vicinity of the point of intersection of the circles $\omega = 0.25$ and $\omega = 0.2$. The actual semicircle has been drawn with some margin of safety. The frequency scale for a point on the semicircle then is chosen, in this case for $\omega = 0.25$. The following is obtained for the function $F_1 F_{c2}$ reckoned from the point +1 for $\omega = 0.25$

$$|F_1 F_{c2}| = 6.3$$

and

$$\varphi = 49.5 \text{ deg}$$

By substitution in Equation [13], $\delta_t = 5$ per cent and $\tau_e = 4.7$ sec are obtained.

The calculated Φ_k-curve obtained with this feedback is shown in Fig. 8. The corresponding efficiency curves together with curves of regulating effort are shown in Fig. 7 (curves 2).

For the sake of comparison, a corresponding calculation has been made with even greater damping by taking $\eta_{r\,min} = -0.04$ as a starting point. The feedback constants obtained were $\delta_t = 16.5$ per cent and $\tau_e = 6$ sec. The corresponding curves for η_r and β_r are shown in Fig. 7 (curves 3).

In the preceding construction only the minimum value of η_r was assumed. More complicated conditions may be given and the problem solved in a similar manner.

A machine with permanent speed droop can be treated similarly by including the feedback F_{c1} in the first link F_1. With reference to Fig. 3, the following relationship will then be obtained

$$F_{11} = \frac{F_1}{1 + F_1 F_{c1}} = \frac{1}{\delta_p} \frac{1}{1 + j\omega \dfrac{\tau_0}{\delta_p}} \quad [17]$$

The open-loop characteristic becomes

$$\Phi_0' = F_{11} F_a F_\xi F_u \quad [18]$$

and when the transient speed droop is connected

$$\Phi_k = \frac{\Phi_0'}{1 + F_{11} F_{c2}} \quad [19]$$

where, referring to Equations [3] and [17]

$$F_{11} F_{c2} = \frac{\delta_t \tau_c}{\delta_p} \frac{j\omega}{1 + j\omega \left(\dfrac{\tau_0}{\delta_p} + \tau_e \right) + (j\omega)^2 \dfrac{\tau_0 \tau_e}{\delta_p}} \quad [20]$$

The corresponding curves for the efficiency and the regulating effort are shown in Fig. 9 for a permanent droop of 2 per cent and the following damping: Curves 1, $\delta_t = 0$; curves 2, $\delta_t = 3.2$ per cent and $\tau_e = 10$ sec.

Network Regulation

In practice, several machines will share the regulation in the network, as shown by the block diagram in Fig. 10. It will be realized that the open-loop transfer function will become

$$\Phi_{kt} = F_a F_{T1} + F_a F_{T2} + F_a F_{T3} + \ldots \text{etc} \quad [21]$$

the summation being vectorial.

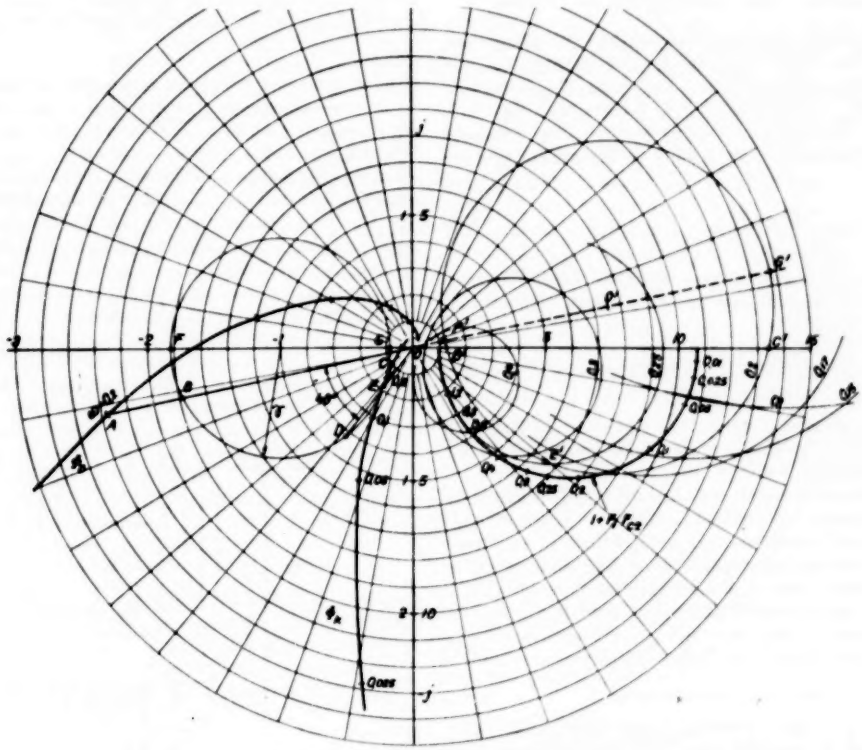


Fig. 8 Graphical Determination of Feedback Constants

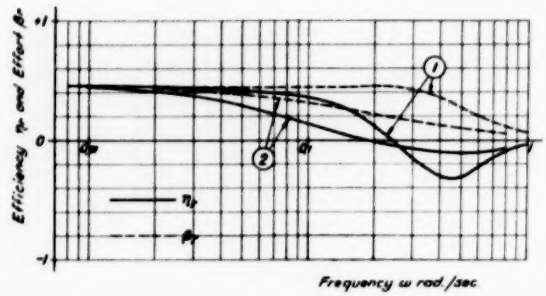


FIG. 9 EFFICIENCY AND REGULATING EFFORT FOR ONE MACHINE HAVING 2 PER CENT PERMANENT SPEED DROOP

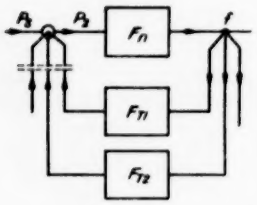


FIG. 10 BLOCK DIAGRAM WITH SEVERAL REGULATING MACHINES

The efficiency for the whole system may be computed as shown previously. Fig. 11 shows the result with a master station with two 50-mw machines and a supporting station with three machines of the same kind. The feedback constants were, respectively

$\delta_p = 0$ per cent $\quad \delta_t = 16.5$ per cent $\quad \tau_e = 6$ sec
$\delta_p = 2$ per cent $\quad \delta_t = 1.25$ per cent $\quad \tau_e = 4$ sec

It should be noted here that the master station, to conform with the original scheme, should have much heavier damping.

It is also possible to calculate the contribution of each machine toward the total regulating effort. It follows from the block diagram in Fig. 12 that the frequency response of the open loop Φ_{kn} for the solitary machine, with all the other machines in circuit, is

$$\Phi_{kn} = \frac{P_{Tns}}{P_{s2}} = \frac{F_n F_{Tn}}{1 + \Phi_{kt} - F_n F_{Tn}} \quad [22]$$

The regulating effort being defined as the relationship between the static power (change in gate opening) and the disturbance, the following is obtained

$$\beta_{rn} = \left|\frac{P_{Tns}}{P_s}\right| = \left|\frac{1}{F_\xi} \frac{\Phi_{kn}}{1 + \Phi_{kn}}\right| \quad [23]$$

By evaluation of these equations, the curves η_r, β_{rk} (unit of a master station), and β_{rs} (unit of a supporting station), as shown in Fig. 11, are obtained.

These over-all curves have been obtained in this example by using only one kind of machine for all stations. It will be realized that the actual frequency range could have been covered more efficiently and with less regulating effort if stations with appropriate characteristics were used for the different ranges of frequency. Thus steam-power stations could be used for quick disturbances of short duration and water-power stations with long water-starting times for slow but sustained changes in the load.

NETWORK DISTURBANCES

The next step is to characterize the actual disturbances occuring in the system by resolving them into their component frequencies, and thus to obtain a direct connection between the disturbances and the regulating properties. In practice, this will be performed by observing the frequency variations in the network which are caused by the disturbances. A characteristic recording is shown in Fig. 13.

In most cases it will not be necessary to carry out a complete analysis of the curve, but will be sufficient to evaluate the amplitudes and frequencies of the predominant disturbances.

Any such evaluation can be made if the relationship is known between the recorded frequency variations and the magnitude of the power disturbances causing them. It is assumed for the sake of discussion that the recording shown holds for a system such as the one chosen as an example. The relationship between the amplitude of the disturbance P_s and the frequency disturbance f it causes is obtained as follows:

According to Fig. 6, for a disturbance of frequency ω_n we have

$$|P_2|\,|F_n| = |f| \quad [24]$$

and by definition, Equations [7] and [8]

$$|P_2| = |P_s|\,(1 - \eta_r) \quad [25]$$

thus

$$|P_s| = \frac{|f|}{|F_n|(1 - \eta_r)} \quad [26]$$

The curve for P_s against ω is shown in Fig. 14 for $f = 0.1$ c/s.

The procedure for estimating the size of the supporting station thus would be as follows:

The amplitude and frequency of the predominating disturbance is estimated from the recording of the frequency variations. Fig. 14 then is used as an aid in determining the amplitude of the corresponding power disturbance. The additional station must be capable of regulating at the actual frequency of the disturbance (i.e., efficiency positive and sufficiently large), and it must

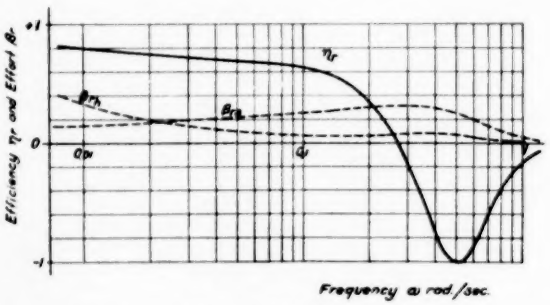


FIG. 11 REGULATING EFFICIENCY IN POWER SYSTEM AND REGULATING EFFORT OF MACHINES

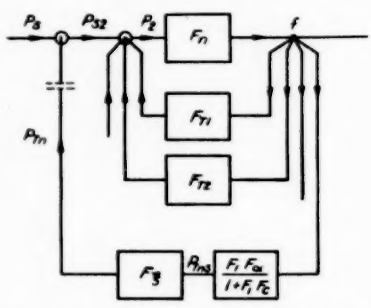


FIG. 12 BLOCK DIAGRAM WITH SEVERAL REGULATING MACHINES. VIEWED FROM ONE OF MACHINES

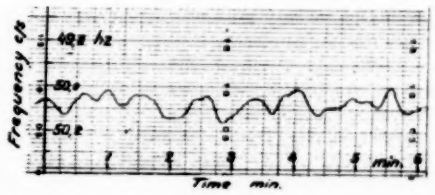


FIG. 13 FREQUENCY VARIATIONS IN REGULATED NETWORK

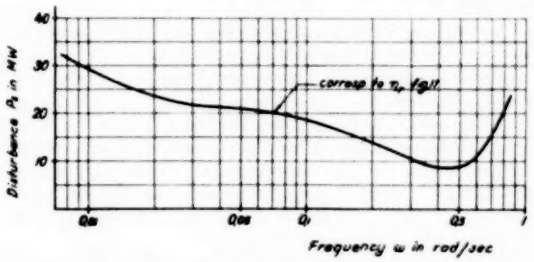


FIG. 14 AMPLITUDE OF DISTURBANCE GIVING 0.1 CYCLE PER SEC FREQUENCY DEVIATION IN REGULATED NETWORK

be capable of supplying sufficient power to eliminate the disturbance. After this estimate, the calculation must be repeated with the new station connected in parallel with the others.

CONCLUSION

It will appear from the foregoing that the regulating activity of each power station in the network should be brought into accord with its ability to be regulated by the use of appropriate damping. An exaggeratedly low damping will give greater activity without improving the regulation in the same degree. Finally, the number and characteristics of stations to be used for regulation will depend on the amount of disturbances in the network.

It also can be seen that the sensitivity of the governors has a minor influence, as long as their action is linear and the gain sufficiently high to make the feedback predominant. Other feedback characteristics than the classical, i.e., proportional to the gate speed, have not been considered in this paper.

Admittedly, the described calculation involves many approximations. However, as the regulation is either slowed down or else carried out with small amplitudes around a mean value, these errors will be kept within reasonable limits. Because of the variation of the network characteristics with the time, a rigorous precision cannot be expected in any case.

Even though the described method of calculation has not been used in the Swedish network to an extent that would give practical experience data, it is expected to give good data for planning the system regulation that will result in an adequate over-all performance at minimum effort. The method also makes it possible to adapt regulation to changes in the working conditions of the power system.

ACKNOWLEDGMENTS

Acknowledgment is given to ASEA, Sweden, under whose auspices this work has been performed and who have permitted the publication of this paper. Acknowledgment is also made to the Swedish State Power Board, Vattenfallsstyrelsen, for sponsoring and collaborating on this development.

BIBLIOGRAPHY

1 "The Swedish Network, Actual and Planned Regulation System," by B. Lof, Report No. 315, part 2 of the CIGRE Meeting, Paris, France, 1950.

2 "An Electrohydraulic Governor for Water Turbines," by S. E. Hedstrom, *ASEA Journal*, vol 25, 1952, pp. 44–56.

3 "Electrohydraulic Control of Water Turbines," by A. Garde, Transactions of the Conference on Automatic Control, Cranfield, England, 1951; *Automatic and Manual Control*, Butterworth Scientific Publications, London, England, 1952.

4 "Investigations by Frequency Analysis of the Swedish Network," by K. Almstrom and A. Garde, Report No. 315, part 3, of the CIGRE Meeting, Paris, France, 1950.

5 "A Method for Calculating the Frequency Regulation Properties of Power Stations," by V. Oja, *ASEA Journal*, vol. 25, 1952, pp. 57–66.

6 "Regulators and Regulator Problems in the Harspranget Power Plant," by S. E. Hedstrom, K. E. Johansson, G. Linden, and B. Olsson, *ASEA Journal*, vol. 26, 1953, pp. 52–64.

Discussion

H. M. PAYNTER.[3] The writer is gratified to see this paper included among those of the Frequency Response Symposium especially since his own "baptism of fire," so to speak, was in this particular battleground of the regulation and control craft.

The problem of power-system governing and load-frequency control is marked by several features sharply in contrast with typical industrial process-control situations or servomechanism problems; some of these perhaps might be summarized in the following:

1 The problem is characterized by an inherent complexity as a result of the great number of generally disparate machines operating with a multiplicity of parallel feedback paths and interconnected by many flexible, electrical links, with numerous essential nonlinear features occurring throughout the problem.

2 The stability and optimum performance of individual machines often must be distinguished clearly from the stability and optimum performance of the entire operating network, both for economic and for technical reasons.

3 The techniques required for successful analysis in this area span many fields of engineering—hydraulic, mechanical, electrical, and so on—and embrace most physical media, including fluid, thermal, electromagnetic, elastic, and the like, thus calling for a combination of sound logical rationale together with practical insight.

4 Significant experiments on the coupled system under normal operating conditions can only be carried to a limited extent.

5 For the reasons just cited, as well as others, practical analytical accomplishment in this area must employ to the maximum extent "model" approaches and other approximation procedures, together with fullest utilization of the improved computing tools developed largely within the past decade.

The author has demonstrated many of the foregoing points in an exemplary fashion. The writer, however, wishes to call attention to signal contributions which have been made by other writers in this field. For the dynamics of hydroelectric plants, particularly with regard to the frequency-response characteristics, the work of Gaden (7),[4] Evangelisti (8), Cuenod (9), and Daniel (10) is of particular note, in addition to many other sources listed in the AIEE "Bibliography on Feedback Control" (11). The writer has discussed specifically some of the hydraulic and governing problems treating frequency-response aspects as well as some of the more important nonlinear elements (12, 13).

Related analytical work dealing with the transient performance of interconnected power systems is now in progress both in this

[3] Assistant Professor of Hydraulic Engineering, Massachusetts Institute of Technology, Cambridge, Mass. Mem. ASME.

[4] Numbers in parentheses (7 to 15, inclusive) refer to the Bibliography at the end of this discussion.

country and abroad; specific attention may be called to the recent papers of Concordia and Kirchmayer together with accompanying discussions by the writer and others (14, 15).

These last cited references feature the use of analog computing equipment for such studies.

Perhaps the motive for the writer's remarks here may be attributed to the long time it has taken to arouse the various responsible parts of the engineering profession to the point where joint and sustained attacks on these problems can be made. He believes that the present paper is another rallying post in this cause.

BIBLIOGRAPHY

7 "Contribution à l'Étude des Regulateurs de Vitesse. Considerations sur le Problème de la Stabilité," by D. Gaden, Éditions la Concorde, Lausanne, Switzerland, 1954.

8 "La Regolazione Delle Turbine Idrauliche," by G. Evangelisti, Zanichelli, Bologna, Italy, 1947.

9 "Influence des Phènoménes de Coup de Bèlier sur le Reglage de la Vitesse des Turbines Hydrauliques," by M. Cuenod, *La Houille Blanche*, Grenoble, France, Mars-Avril, 1949, pp. 163–182.

10 "Accélération du Reglage de Vitesse des Turbines Hydrauliques," by Jean Daniel, *La Houille Blanche*, nos. 1 and 2, 1948.

11 Bibliography on Feedback Control—Part II, AIEE Technical Paper 53-251, 1953; see particularly section T: "Prime-Mover Speed Regulation," and section U: "Power System Regulation."

12 "Electrical Analogies and Electronic Computers: Surge and Water Hammer Problems," by H. M. Paynter, Trans. ASCE, vol. 118, 1953, p. 962.

13 "Methods and Results From M.I.T. Studies in Unsteady Flow," by H. M. Paynter, *Journal, Boston Society of Civil Engineers*, vol. 39, no. 2, April, 1952.

14 "Tie-Line Power and Frequency Control of Electric Power Systems," by C. Concordia and L. K. Kirchmayer, Trans. AIEE, 1953, "Power Apparatus and Systems," p. 562.

15 "Tie-Line Power and Frequency Control of Electric Power Systems—Part II," by C. Concordia and L. K. Kirchmayer, AIEE Technical Paper 54-25.

F. R. Schleif.[8] The author has made a valuable and significant contribution to the technical literature in the somewhat slighted field of power system frequency regulation. His analysis by the method of frequency response has evaluated the important factors to bring large power system frequency regulation within engineering science. It marks considerable contrast with the all too recent appraisal that only two requirements must be met; the system must be stable and deviations must not be so bad as to cause customer complaint. His analysis is further an advance over the experimental tailoring of regulating characteristics of large generating units to the system into which they feed by means of Léauté diagrams relating frequency deviations with gate movement.

It is particularly significant that, in the analysis the author describes, the regulating efficiency of individual units under various frequencies of disturbance is determined, and useless regulating activity, wherein the response is out of phase with the disturbance, may be limited. The desirable combination of steam units for regulating the small fast variations and hydrounits for the larger slow variations becomes evident.

In the experimental determination of the system transfer function F_s it is not stated, but we assume that the change in the function with change in load conditions has been influenced by the change in system inertia by removal of generating units as the load drops. Is this correct? Since some change of inertia must accompany change of load, it seems appropriate also to include whatever change due to dispatching of generation would be typical of operation of this particular system.

J. H. Drake.[9] The recent published work of the Swedish engineers in the field of turbine regulation has been followed with considerable interest by Southern California Edison Company engineers. The Edison and Woodward Governor Companies are embarking on a co-operative program which has the following major aspects:

1 Development of suitable instrumentation and test equipment for the routine adjustment of governors, as well as the development of governor test methods.

2 Determination of optimum governor settings based on the dynamic behavior of the primary elements of the over-all feedback loop.

The program is in the beginning stages and is now pending development of system instrumentation.

We would appreciate it if the author would answer and expand on the following:

1 It is stated that Equation [5] closely follows the Nyquist plot, particularly at low values of the forcing-function frequency. It would appear that this expression, which is equivalent to the admittance of a series R, L electric circuit, could be obtained by subjecting a power system to an impact load and making an analysis of a high-speed frequency record.

Did the author find that it was important to know the manner in which the Nyquist plot departs from a semicircle at higher forcing-function frequencies theoretically to determine optimum governing effort?

2 Were Nyquist plots obtained at more than one point in the Swedish network to verify the validity of assuming that "the simple expression [5] may be used for preliminary analysis at all significant parts of the network?"

3 The author apparently attributes the dominant period of the slow system-frequency variation in Fig. 13 to be determined by the frequency of the predominating power disturbance. Has it been determined which load or loads cause the predominating power disturbance?

Unless all governors were in a blocked condition for the time period represented in Fig. 13, it would appear that the dominant single low-frequency variation is possibly caused by the combined oscillation of all machines and their governors in response to load variations of a random nature, rather than by load variation of a cyclical nature.

The author and the other Swedish engineers who contributed to material presented in the paper are to be commended for the original analysis displayed in attacking a complex problem which heretofore has been approached with "cut-and-try" procedures. The engineers of the Southern California Edison and Woodward Governor Companies are anticipating that after system tests are performed a theoretical analysis of power-system regulation can be made and subjected to verification in actual practice.

R. Oldenburger.[7] The Swedish power system is probably the largest physical system on which frequency-response runs have been made. This is particularly significant in view of the tremendous number of pieces of equipment that make up such a system. The fact that the simple transfer function of Equation [5] of the paper fits the experimental data so well implies that the entire system behaves as if it were a single machine with a first-order lag.

The writer's company is working with the Southern California Edison Company in collaboration with the General Electric Company on making such runs on the Southern California Edison power system. From the distribution of generating units and loads it appears that this system must be treated as a three-machine system, one at Hoover Dam, another at Big Creek, and a third at Los Angeles.

[8] Engineer, U. S. Bureau of Reclamation, Denver, Colo.

[9] Engineer, Southern California Edison Company, Ltd., Los Angeles, Calif.

[7] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

The question of the general validity of Equation [5] is an important and serious one. The value of frequency-response runs is much enhanced if the differential equations of the system on which the runs are made are worked out in advance, and these runs are made to verify the theory. The problem of working out such equations for a power system is one of major proportions.

To properly design governors and other controls for power systems the analyst must have the differential equations for these systems or equivalent information. The controls obviously should fit the specific system to give optimum regulation at minimum cost.

In place of the efficiency introduced by the author in terms of power, one might use the deviation ratio, discussed by Janssen in his Frequency Response Symposium paper No. 53—A-21, and apply it to the frequency of the power system; that is, use the deviation of the frequency with governing divided by this deviation without governing. After all, one is interested in controlling this frequency.

P. S. Creager.[8] The author has written a very interesting paper. The writer was especially impressed by the ingenious scheme of integrating power output to eliminate random load effects.

The technique used here appears to be quite novel and to have a large number of applications. The writer would suggest that the author present this method in detail in a future paper. If this does not appear practicable, the writer would ask him to include such a discussion in his formal closure, so that it will appear in permanent form.

The writer is not clear about a few points. Under the heading, Transfer Function of the Network, in the first paragraph, we find "The output of one 50-mw machine was varied sinusoidally at various frequencies." From the presentation of the paper, the writer understood that the sine variation actually was impressed on the electronic portion of the governor device.

Does this mean then (*a*) that the ratio of power as measured by the integrating scheme to the voltage impressed on the governor will give the transfer function of boxes F_1, F_a, and F_t, of Fig. 3, and (*b*) that the ratio of the resultant frequency deviation to this power gives the data plotted in Fig. 5?

Was this frequency deviation taken from a record such as Fig. 13? What was the amplitude of the sinusoidal power variation?

Author's Closure

In order to give in a limited space an over-all picture of the development of frequency regulation in the Swedish network, some concentration was necessary and treatment of the more general application of the described calculations was not undertaken. This question has been brought up through the discussions and will be very shortly dealt with here. In order to give more complete answers, some of the questions have been referred to the Bibliography.

First, the derivation of Equation [4], the influence of water inertia, including a list of assumptions, will be found in reference (3) of the Bibliography. The most significant limitation to more general use in similar calculations is the assumption of nonelasticity of water and pipe line. This assumption has been adequate for conditions in Sweden, where the significant stations have practically vertical penstocks located in the dam or taken through rock. Subsequently, similar equations have been developed even for other conditions in the work mentioned as reference (15) in Professor Paynter's Bibliography.

Second, the question of a more general validity of the network Equation [5] has been brought up in several discussions. This takes us back to the question of whether the inertia effects of the more or less rigidly interconnected machines (both synchronous and asynchronous) in a power system can be lumped together into one group, or whether they would have to be divided into two, three, or even more rigid groups interconnected by flexible links. In general terms, the number of groups to be used for a study of governor behavior will depend on the maximum frequency appearing with appreciable amplitude in the disturbing function and for which governor action is really obtained—for which, consequently, reasonable accuracy of the equation is needed—as compared to the natural oscillation frequencies pertinent to the system of several rigid groups with flexible ties. If the natural frequencies of oscillation are high as compared to the maximum frequency for which governor action is obtained, the adopted network model can be further simplified. This very general and known criterion seems to be the most practical since the frequency of oscillations, that are caused by distributed inertias (not by governor action) are, in general, approximately known. In the Swedish network the observed oscillations are of the order of 1 cps which makes it reasonable to assume validity for the simple Equation [5] up to 0.1 cps and even somewhat higher, giving sufficient accuracy for governor studies at normal conditions in the network.

The description of the network measurements including methods employed, results obtained, interpretation of the constants, and their variations at different load conditions will be found in reference (4) of the Bibliography.

The transfer function of the network could most likely also be obtained by subjecting the system to a step-function disturbance and recording the frequency variations. To eliminate the random frequency variations, this recording must be repeated a sufficient number of times and the results superimposed. This method would also have the effect of revealing in a more direct manner the natural frequencies of oscillation inherent to the system.

The frequency recording on Fig. 13 represents frequency variations obtained by the combined action of disturbance and governor movement. The possible accentuation of certain frequencies by the governors is reflected by the curve in Fig. 14 that is used for interpretation of the recorded frequency variations.

Finally, appreciating the constructive discussions, the writer would like to conclude with a general comparison of trends in some more modern interpretations of network problems. The existence of weak tie lines combined with the prevalence of steam power have in this country led to a treatment of power systems as several group structures with special emphasis on tie-line regulation and without differentiating between the various characteristics of the stations within the different groups, the damping being no problem for steam plants and being tied to disconnected no-load operation in hydro plants. In Sweden, the lack of tie-line problems and the early development of the electrohydraulic governor for hydro stations have led the trend toward the organization of co-operation between the different plants in the group itself, leading to single-group treatment with emphasis on the different characteristics of the individual stations. The increasing complexity of power systems through new interconnections, together with the increased use of modern computing machines and a growing insight into the economical advantages of co-ordinated regulation in power systems, is likely to result in a development where both aspects of the general problem will be considered.

[8] Associate Professor of Electrical Engineering, Rutgers University, New Brunswick, N. J. Mem. ASME.

The Use of Zeros and Poles for Frequency Response or Transient Response

By W. R. EVANS,[1] DOWNEY, CALIF.

The characteristic equation of a control system is often of the form $1/KG(s) + 1 = 0$, in which $G(s)$ is $(s - q_1)(s - q_2) \ldots /(s - p_1)(s - p_2) \ldots$. Each factor of $G(s)$, such as $(s - p_1)$, is a complex variable which can be plotted as a vector in the s-plane from the pole p_1 to the point s. The complex product $G(s)$ is itself a vector whose magnitude is the product of magnitudes and whose angle is the sum of the angles of its factors. Frequency response is computed directly for s points along the $j\omega$-axis. In order to find the roots of $1/KG(s) = -1$, the locus of s is sketched, where for each point on this locus the total angle of $G(s)$ is equal to 180 deg; a point on this locus is then a root if $K = 1/|G(s)|$. For given initial conditions the Laplace transform permits the transient response of the system to be written directly.

Frequency Response From Zeros and Poles

THE servo shown in Fig. 1 is used to illustrate all of the calculations involving zeros and poles for frequency response or transient response. The system is simplified by

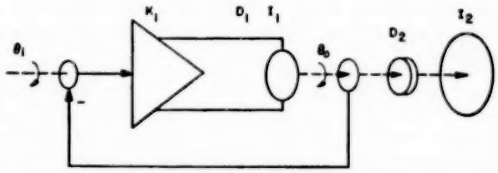


Fig. 1 Servo With Auxiliary Damping

letting $D_2 = 0$, and then the equation for torques on the output shaft is

$$K_1\theta_e = D_1\frac{d\theta_0}{dt} + I_1\frac{d^2\theta_0}{dt_2} \quad \ldots\ldots\ldots\ldots [1]$$

The torque $K_1\theta_e$ is the stall torque of the motor produced by the amplifier voltage which is directly proportional to the error angle, $\theta_e = \theta_i - \theta_0$. The torque $D_1 d\theta_0/dt$ is the decrease in motor torque in proportion to speed; I_1 is the total inertia directly connected to the output shaft.

For the frequency response of the system only sinusoidal signals need be considered. As is well known, a sine wave can be represented as the imaginary part or the projection on the vertical axis of a rotating vector $e^{j\omega t}$. Let $\theta_e = \Theta_e e^{j\omega t}$, in which Θ_e in general, is a vector $|\Theta_e|e^{j\phi_e}$ with an initial phase angle ϕ_e. Substituting for θ_e and θ_0 gives

$$K_1\Theta_e e^{j\omega t} = D_1 j\omega\Theta_0 e^{j\omega t} + I_1(j\omega)^2\Theta_0 e^{j\omega t} \quad \ldots\ldots [2]$$

Note that the factor $j\omega$ arises for each differentiation, and that the unit rotating vector $e^{j\omega t}$ remains in all terms. Dividing through by $e^{j\omega t}$ and solving for the ratio of the Θ_e and Θ_0 vectors gives

$$\frac{\Theta_e}{\Theta_0} = \frac{D_1 j\omega + I_1(j\omega)^2}{K_1} = \frac{j\omega(j\omega + D_1/I_1)}{K_1/I_1} \quad \ldots\ldots [3]$$

This ratio can be calculated for various values of ω from the vector plot in Fig. 2(*a*) as the heads of both vectors move up the $j\omega$-axis and is shown in Fig. 2(*b*). For a given choice of the constant K_1 the scale of the drawing is fixed, and the ratio of Θ_i/Θ_0 is equal to $(\Theta_0 + \Theta_e)/\Theta_0$, or $1 + \Theta_e/\Theta_0$. This vector sum is thereby obtained as a vector to the same curve with the base at the -1 point.

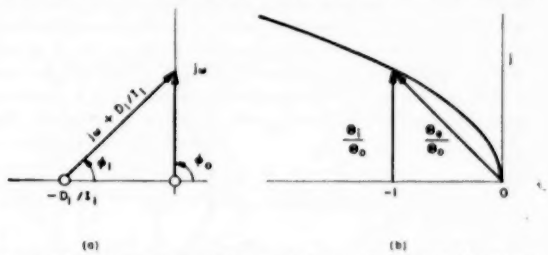


Fig. 2 Calculation of $\Theta_e/\Theta_0 = j\omega(j\omega + D_1/I_1)/(K_1/I_1)$ From Vectors in (*a*) and Plotted in (*b*)

In the more general case of the transient plus steady-state solution of the differential Equation [1], damped sine waves are involved. The damping factor $e^{\sigma t}$, in which σ is a negative number for a decay, can be combined with $e^{j\omega t}$, as in

$$\theta_0 = \Theta_0 e^{\sigma t} e^{j\omega t} = \Theta_0 e^{(\sigma + j\omega)t} = \Theta_0 e^{st} \quad \ldots\ldots [4]$$

Thus s is, in general, a complex variable which would appear in place of $j\omega$ in Equation [2] or [3]. The general expression for the input to output ratio is then

$$\frac{\Theta_i}{\Theta_0} = 1 + \frac{\Theta_e}{\Theta_0} = 1 + \frac{s(s + D_1/I_1)}{K_1/I_1} \quad \ldots\ldots [5]$$

The characteristic equation of the system is Equation [5] for the special case of the input Θ_i being zero. The roots of this equation are the values of s which make $\Theta_e/\Theta_0 = -1$. This function can be calculated by a slight modification of Fig. 2(*a*) in which s is a complex number rather than simply $j\omega$, as in Fig. 3(*a*). The complex product of the vectors must have a total phase angle of 180 deg. The product locus is a line perpendicular to the real axis bisecting the line segment between the zeros. This result can be checked by observing that the quadratic formula gives as the two roots

$$s_{1,2} = \frac{-D_1}{2I_1} \pm j\sqrt{\frac{K_1}{I_1} - \left(\frac{D_1}{2I_1}\right)^2} = \sigma \pm j\omega \quad \ldots\ldots [6]$$

If the root is desired at the point s_1 shown, then the magnitude K_1/I_1 must equal the complex product of the vector lengths, in

---

[1] Systems Group Leader, Electro-Mechanical Engineering Department, North American Aviation, Inc.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29-December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 11, 1953. Paper No. 53—A-24.

order to make $|\Theta_e/\Theta_0| = 1$. Knowing the roots permits the function in Equation [5] to be rewritten as

$$\frac{\Theta_i}{\Theta_0} = (1 - s/s_1)(1 - s/s_2) = \left(\frac{s - s_1}{-s_1}\right)\left(\frac{s - s_2}{-s_2}\right) \quad \ldots [7]$$

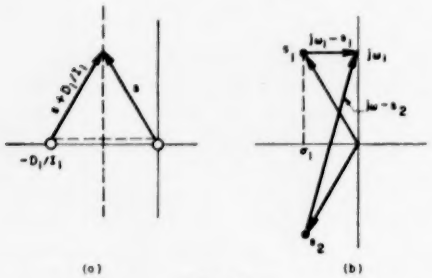


FIG. 3
(*a*, Root locus versus K_1 for $s[s + D_1/I_1]/[K_1/I_1] = -1$. *b*, Frequency response of $[1 - s/s_1][1 - s/s_2]$.)

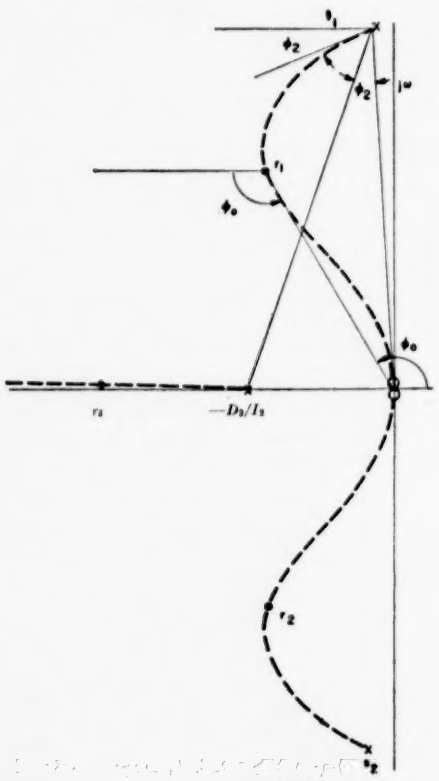


FIG. 4 ROOT LOCUS OF $\dfrac{D_2s^2/K_1}{(1 - s/s_1)(1 - s/s_2)(s + D_2/I_2)}$ WITH $D_2/I_2 = 0.4\ |s_1|$; ROOTS FOR $D_2/K_1 = 1/|s_1|$

The frequency response can be calculated directly from Equation [7] using the vectors in Fig. 3(*b*). Note that each term such as $(1 - s/s_1)$ is the complex ratio of two vectors, $(j\omega - s_1)/(-s_1)$. The minimum value of this ratio would occur for $\omega = \omega_1$ and be equal to σ_1/s_1. This ratio is the damping ratio of the root ζ which also can be specified as cosine of ψ_0. Note that the minimum value of Equation [7] is determined strictly by the damping ratio of the root. Similarly, the decay of a transient per cycle of oscillation $e^{\sigma(2\pi/\omega)}$ depends only on the damping ratio.

LOAD-DAMPED SERVO

In many servos the required torque per unit error angle K_1 is sufficiently great so that the damping ratio which would result from motor characteristics alone is insufficient. Greater damping can be achieved by addition of a dashpot between the output shaft and a fixed base, but this would have the undesirable effect of increasing the error required to drive the output at a fixed velocity. If a dashpot is connected to an auxiliary mass, however, the mass eventually would get up to the same speed as the output shaft and thereby no longer apply a drag torque. The equation for torques on the output shaft of the complete system is therefore

$$K_1(\Theta_i - \Theta_0) = D_1s\,\Theta_0 + I_1s^2\,\Theta_0 + D_2s(\Theta_0 - \Theta_d) \ldots [8]$$

The dashpot torque $D_2s(\theta_0 - \theta_d)$ is equal to $I_2s^2\theta_d$. Solving for Θ_d, substituting into Equation [8], and solving for the ratio of Θ_i/Θ_0 gives

$$\frac{\Theta_i}{\Theta_0} = \frac{K_1 + D_1s + I_1s^2}{K_1} + \frac{D_2s}{K_1}\left(1 - \frac{D_2s}{D_2s + I_2s^2}\right) \ldots . [9]$$

The first term in this function had its zeros s_1 and s_2 already determined; the quantity in parentheses can be simplified further to give

$$\frac{\Theta_i}{\Theta_0} = (1 - s/s_1)(1 - s/s_2) + \frac{D_2s}{K_1}\left(\frac{s}{s + D_2/I_2}\right) \ldots [10]$$

The characteristic equation is Equation [10] for $\Theta_i = 0$, and the equation is satisfied if the ratio of the second term to the first term is -1

$$\frac{D_2s/K_1}{(1 - s/s_1)(1 - s/s_2)}\left(\frac{s}{s + D_2/I_2}\right) = -1 \ldots \ldots [11]$$

For I_2 a very large simplified root-locus plot results. In it at least one root is always close to the origin. Therefore the transient response would have a term with a long decay time. This result would be expected physically because the drag torque on the output shaft for a change in input velocity would not be eliminated until I_2 was accelerated.

The root close to the origin can be avoided by moving the pole $(-D_2/I_2)$ out along the real axis as in Fig. 4. The initial direction of the locus from s_1 will be found to be nearly ϕ_2 below the horizontal by finding the total angle of all vectors at s_1 except $(s - s_1)$ in Equation [11] and then solving for the angle of $(s - s_1)$ needed to bring the total angle to 180 deg. The rest of the locus is sketched to approach the zeros at the origin along the j-axis. The locus at r_1 is checked by adding all of the vector angles at this point. Note that any angle such as ϕ_0 is present at the point r_1 being checked. A transparent sheet pivoted at r_1 could be rotated successively through each of the angles in order to obtain the sum. One or two trials are usually sufficient to locate the locus in the region of interest. In this case, r_1 is then selected for maximum damping. The vector lengths to the point r_1 are multiplied as indicated in Equation [11] and D_2/K_1 assigned the value such as to make the magnitude of the function in Equation [11] equal to 1. This calculation can be aided by a scale pivoted at the r_1-point and marked with the logarithm of the distance from the pivot point. Addition of logarithms then gives the logarithm of the product. For the given plot in which $D_2/I_2 = 0.4\ |s_1|$, $D_2/K_1 \cong 1/|s_1|$ for $s = r_1$. With value of D_2/K_1 selected, the root r_2 would, in general, have to be located by trial and error, but in this case it can be calculated directly by comparison of the coefficient of the s^2 term in Equation [10] with the

s^3 term for the same function written in terms of the roots

$$\frac{\Theta_i}{\Theta_0} = \frac{(1 - s/r_1)\,(1 - s/r_2)\,(1 - s/r_3)}{(1 + I_2 s/D_2)} \quad \ldots\ldots [12]$$

The function in Equation [10] would have to be put over a common denominator. Then equating coefficients of the s^3-term permits direct solution for r_2. It is $(-D_2/I_2)\,(|s_1|^2/|r_1|^2)$. Incidentally, comparison of coefficients reveals that any system having two or more poles than zeros has a constant sum of roots. The frequency response of Equation [12] can be calculated from vectors, as indicated in Fig. 5(*a*). At $s = j\omega_2$ the minimum introduced by the pair of complex zeros is 0.90. The terms involving

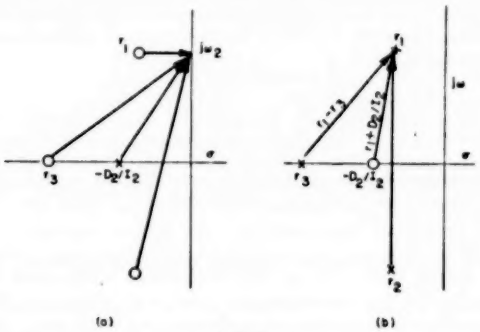


FIG. 5
(*a*, Vector for frequency response. *b*, Vectors for amplitude of transient term.)

the zero and pole along the real axis, however, have a value of 0.70 at $s = j\omega_2$, which makes the approximate minimum of the complete function be 0.9×0.7, or 0.63. The corresponding resonant peak of Θ_0/Θ_i would be 1.6, corresponding to a damping ratio for a simple quadratic system of 0.33, compared to 0.50 for the roots themselves. This example shows that the estimate of damping ratio of a simple quadratic from the resonant peak of the frequency response is not always valid in estimating the transient response of the system.

TRANSIENT RESPONSE

The algebraic equations used thus far could have been obtained formally by use of the Laplace transform. The transform has the advantage of permitting direct solution for the amplitude of transient terms, thus avoiding the classical method of setting up a general expression for the transient and then specifying the initial conditions which determine amplitudes from the solution of simultaneous equations. For a step-velocity input the "transform" of θ_i is $\dot{\theta}_i/s^2$. The transform of the error is equal to the transform of the input times the transfer function (Θ_e/Θ_i) for the special case of all initial conditions being zero

$$\Theta_e(s) = \Theta_i(s)\,\frac{\Theta_e}{\Theta_i}(s) = \Theta_i(s)\left[1 - \frac{\Theta_o}{\Theta_i}(s)\right]$$
$$= \frac{\dot{\theta}_i}{s^2}\left[1 - \frac{1 + I_2 s/D_2}{(1 - s/r_1)\,(1 - s/r_2)\,(1 - s/r_3)}\right] \quad \ldots [13]$$

The error as a function of time is the sum of exponential terms of the form $A_i e^{s_i t}$, in which one term arises for each pole of Equation [13]. The case of duplicate poles, such as exist in the origin in this case, can be handled by assuming a slight separation of the poles, expressing the function of time, and then letting the separation of poles approach zero. In this case, however, the constant and velocity term which normally would arise due to s^2 are zero if the constant $D_1 = 0$ and equal to $D_1\dot{\theta}_i/K_1$ otherwise. The amplitude A_1 is found by canceling the factor $(s - r_1)$ in the denominator and letting $s = r_1$ in Equation [13]

$$A_1 = \left[\frac{-\dot{\theta}_i}{s^2}\,\frac{(1 + I_2 s/D_2)}{(-1/r_1)\,(1 - s/r_2)\,(1 - s/r_3)}\right]_{s = r_1}$$
$$\cong \frac{1}{0.70\ \angle 120^\circ}\,\frac{1.6\ \angle 85^\circ}{1.73\ \angle 30^\circ}\,\frac{}{0.95\ \angle 50^\circ} \cong 1.40\ \angle -115^\circ \quad \ldots [14]$$

The root r_2 sets up a term $A_2 e^{r_2 t}$, which is at all times the conjugate of $A_1 e^{r_1 t}$, and therefore the sum of these two terms has the j-part canceled and a real part of twice the magnitude of either. With the vectors in Equation [14] measured from Fig. 5(*b*) the transient response for $\dot{\theta}_i = 1$, $|s_1| = 1$ is given by

$$\theta_e \cong 2.8\,e^{-0.35t}\cos(0.60t - 115^\circ) + 1.1\,e^{-0.35t} \quad \ldots [15]$$

The plot of this function in Fig. 6 is sketched from a few significant points. Thus the time constant of decay of the sine wave $+1/0.35$ permits the envelope to be sketched, the initial phase angle and frequency determine the points where the cosine wave goes through zero, and the wave is sketched to be tangent to the envelope at the mid-point between zero points. In this case the simple exponential $e^{-0.3t}$ dies out rapidly and has little net effect. Frequently a transient can be sketched by establishing only a single damped sine wave and then sketching the first part such as to satisfy initial conditions.

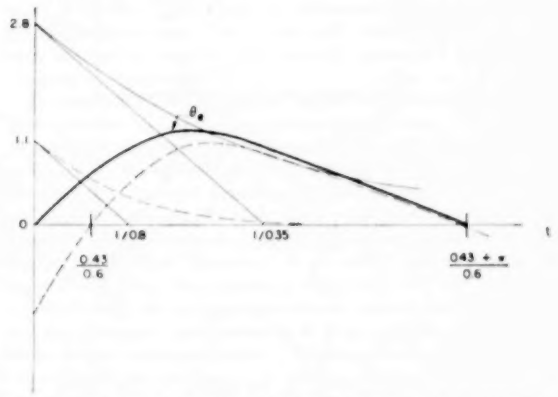


FIG. 6 PLOT OF $\theta_e \cong 2.8\,e^{-0.35t}\cos(0.60t - 115^\circ) + 1.1\,e^{-0.35t}$

TRANSFER FUNCTION BY INSPECTION

One of the advantages of the concept of zeros and poles is the ease of setting up an equation to study the effect of a new parameter. Thus, if the effect of a spring K_2 added between the output shaft and the inertia is to be studied, the characteristic equation of the system can be written in the general form

$$K_2\left[\frac{(1 - s/q_1)\,(1 - s/q_2)\,\ldots\ldots}{Bs^{\theta}(1 - s/p_1)\,(1 - s/p_2)\,\ldots}\right] = -1 \quad \ldots [16]$$

This equation corresponds to setting up a complete block diagram and then treating the entire system as seen from the K_2 loop as a single transfer function, as expressed in the brackets. For K_2 very small, the equation can be satisfied only by s nearly equal to the poles of the function in the brackets. In the limit for $K_2 = 0$, the roots must be the poles. Similarly, for $K_2 = \infty$, the roots must be the zeros. But physically, $K_2 = 0$ makes the system identical with that just solved for which the roots are r_1, r_2, and r_3. Also, for $K_2 = \infty$ the auxiliary inertia is directly added to the output inertia, so that the roots s_i in Fig. 4 would be

brought to a lower natural frequency, as shown by q_1 and q_2 in Fig. 7. Specifically, q_1 can be shown to lie along a circular arc which passes through s_1 and the origin. Note that all terms thus far discussed are equal to 1 when $s = 0$. The function in the brackets has the significance of being the compliance of the rest of the system as seen from the spring. Thus if a constant torque were impressed between the output shaft and the inertia I_2, the steady relative motion which would result would be limited only by the dashpot, whose compliance is $1/D_3$. The general function first expressed in Equation [16] thus becomes completely specific

$$K_3 \frac{(1 - s/q_1)(1 - s/q_2)}{D_{23}s(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)} = -1 \quad \ldots\ldots [17]$$

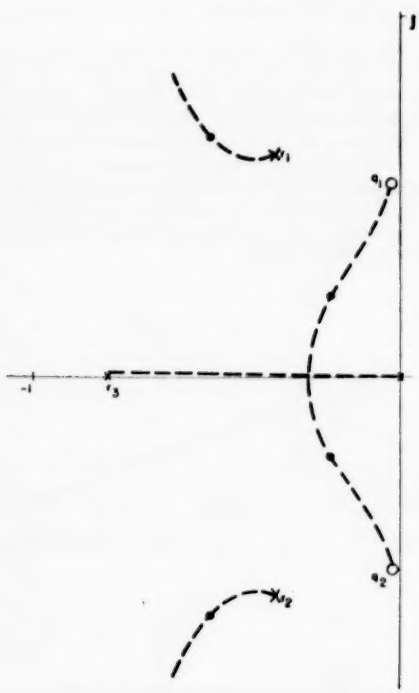


FIG. 7 ROOT LOCUS OF $\dfrac{K_3(1 - s/q_1)(1 - s/q_2)}{D_{23}s(1 - s/r_1)(1 - s/r_2)(1 - s/r_3)} = -1$ ROOTS AT $K_3/D_3 = 0.2\ |s_1|$

The corresponding plot with K_3 variable in Fig. 7 shows that the root originating at r_1 has increased damping as the first effect of the spring. As the spring constant is increased, however, new low-frequency roots develop from poles on the real axis. A choice of $K_3/D_3 = 0.20\ |s_1|$ would result in the roots being at the points shown by the dots in which adequate damping is present in both modes.

A summary of numerical values may be of interest. The magnitude $|s_1|$ was specified as 1 for Equation [15]; but $|s_1| = K_1/I_1$, and therefore let both $K_1 = 1$ and $I_1 = 1$. The damping ratio of s_1 is 0.05; therefore $D_1 = 0.1$. The amplification D_2/K_1 in Fig. 5 is $1/|s_1|$ making $D_2 = 1$; and D_2/I_2 is 0.4 $|s_1|$, making $I_2 = 2.5$. Finally, $K_3/D_3 = 0.2\ |s_1|$, making $K_3 = 0.2$.

The same technique can be used for any system in which the parameter of interest occurs at only one point in the block diagram or determinant of the system. Thus, for the common problem of selecting the value of the condenser in a lead network of a servo, the special case of $C = 0$ corresponds to an open circuit and $C_2 = \infty$ to a short circuit being present in place of the condenser. These extreme cases correspond merely to a change in amplification around the loop and the roots thus can be located along the locus versus amplification. The characteristic equation is $ZCs = -1$, in which the zeros and poles of the impedance Z are therefore known. The locus is then sketched versus C and the roots located for maximum damping or maximum damping ratio.

Complicated Systems

The foregoing techniques are recommended for comparatively simple systems for which a physical picture can be easily maintained. For complicated systems it is convenient to use determinants. Thus consider a system described by three equations and three variables, for which the characteristic determinant is third order, as shown in

$$\Delta = \begin{vmatrix} F_{11}(s) & F_{12}(s) & F_{13}(s) \\ F_{21}(s) & F_{22}(s) & F_{23}(s) \\ K_{31}G_{31}(s) & K_{32}G_{32}(s) & F_{33}(s) \end{vmatrix} \quad \ldots\ldots\ldots\ldots [18]$$

The F-functions are given, whereas the $KG(s)$-functions are to be selected. Expansion by minors about the third row then gives

$$\Delta = K_{31}G_{31}(s)\, M_{31}(s) - K_{32}G_{32}(s)\, M_{32}(s) + F_{33}(s)\, M_{33}(s) \quad [19]$$

Assume at first that all minors such as $M_{31}(s)$ are quadratic functions of s whose zeros may be determined by the quadratic formula. If K_{32} were first assumed to be zero, then the sum of the two remaining terms could be made zero by making the ratio of the terms equal -1, as expressed in

$$\frac{K_{31}G_{31}(s)\, M_{31}(s)}{F_{33}(s)\, M_{33}(s)} = -1 \quad \ldots\ldots\ldots\ldots [20]$$

If the roots of this equation are denoted as r_1, r_3 . . . , then the complete determinant in Equation [19] can now be expressed as

$$\Delta = \frac{K(1 - s/r_1)(1 - s/r_2)\ldots}{s^\beta(1 - s/p_1)(1 - s/p_2)\ldots} - K_{32}G_{32}(s)\, M_{33}(s) \ldots [21]$$

The poles, such as p_1, are the poles of $G_{31}(s)$, $M_{31}(s)$, $F_{33}(s)$, and $M_{33}(s)$. The value of K and β are determined by comparing Equations [19] and [21] for the special case of s nearly zero. The roots of Equation [21] are then found by a second plot which makes the ratio of the remaining terms equal to $+1$. Note that this form of expansion of the determinant keeps the adjustable functions isolated, so that their effect may be determined conveniently.

If the system is more complicated, then the minors which arise will be of higher than second order. One way of finding roots of a higher-order equation using a root-locus plot is permitted by grouping terms as in

$$\left.\begin{aligned} &a_5s^5 + a_4s^4 + a_3s^3 + a_2s^2 + a_1s + a_0 \\ &= s^3(a_5s^2 + a_4s + a_3) + (a_2s^2 + a_1s + a_0) = 0 \end{aligned}\right\} \ldots [22]$$

The zeros of the two functions in parentheses can be found by the quadratic formula. The sum of the remaining two terms is then made zero by making the ratio of the terms equal to -1.

Conclusions

The concept of zeros and poles of transfer functions permits quick estimates of the frequency response of a system in which the entire picture is based upon vector lengths and angles. The roots of a characteristic equation of the form $1/KG(s) + 1 = 0$ can be sketched readily in the s-plane as a function of K, if the zeros and poles of $G(s)$ are known. The characteristic equation

can be obtained in the required form directly, if the roots of the equation are known for two specific values of K with the cases of $K = 0$ and $K = \infty$ frequently being most easily solved. Root-locus plots are a convenient way to keep track of the zeros or poles of a complete system as additional terms or loops are added; the rest of the advantages of zeros or poles are independent of the means by which these points are located.

REFERENCES

1 "Control System Synthesis by the Root Locus Method," by W. R. Evans, Trans. AIEE, vol. 69, 1950, p. 66.

2 "Transients in Linear Systems," by M. F. Gardner and J. L. Barnes, McGraw-Hill Book Company, Inc., New York, N. Y., 1942.

3 "Damper Stabilized Instrument Servomechanisms," by A. C. Hall, Trans. AIEE, vol. 68, 1949, pp. 299–306.

Discussion

YAOHAN CHU.[1] In the analysis of control problems the transient-response method was first employed. Because of the difficulty encountered in the synthesis of control systems by the transient-response method for systems with higher-order ordinary linear differential equations with constant coefficients, the frequency-response method has been developed[2] and widely used. Since the development of the root-locus method by the author, the transient-response method becomes again a useful tool in the synthesis of control problems. Furthermore, this method brings about another way of viewing the correlation between the transient and frequency responses. Both these two points are clearly shown in this paper. The author is to be commended for his contribution. He is now introducing this method to the mechanical engineers.

The root-locus method has its uniqueness in at least two respects:

(*a*) The concept of plotting all possible roots of the characteristic equation of a feedback-control system on the complex plane.

(*b*) The graphical method to construct the locus.

However, for the case of "complicated systems" which the writer interprets to mean multiloop, linear, feedback-control systems, it is generally difficult to apply this graphical method. An example is used to illustrate this point. Fig. 8 of this discussion shows a control system with three feedback loops. The characteristic determinant[4] Δ of this system[5] is

$$\Delta = \begin{vmatrix} 1 & 0 & G_3 \\ -G_1 & 1 & G_3H_{32} \\ -F_{13} & -G_2 & 1 + G_3H_{33} \end{vmatrix} \quad [23]$$

or

$$\Delta = 1 + G_3H_{33} + G_2G_3H_{32} + G_1G_2G_3 + F_{13}G_3 \quad [24]$$

and the characteristic equation is

$$\Delta = 0 \quad [25]$$

or

$$1 + G_3H_{33} + G_2G_3H_{32} + G_1G_2G_3 + F_{13}G_3 = 0 \quad [26]$$

[2] Scientific Laboratory, Ford Motor Company, Dearborn, Mich.

[3] "The Analysis and Synthesis of Linear Servomechanisms," by A. C. Hall, Technology Press, Massachusetts Institute of Technology, May, 1943.

[4] G, F, and H represent, respectively, direct, feedforward, and feedback transfer functions, all of which are functions of the complex variable s.

[5] "A Generalized Theory of Multi-Loop Linear Automatic Control Systems," by Yaohan Chu, doctorate thesis, Massachusetts Institute of Technology, 1953.

The root-locus of this system for a certain gain parameter is, in general, not feasible to be constructed by the author's graphical method because there is more than one complex term excluding

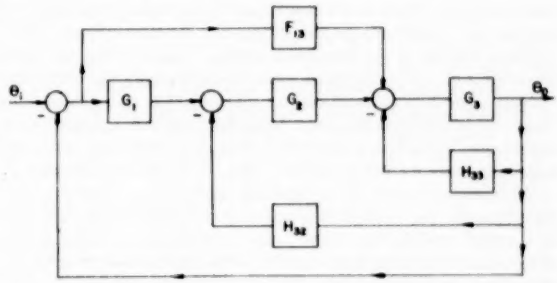


FIG. 8 BLOCK DIAGRAM OF A CONTROL SYSTEM WITH THREE FEEDBACK LOOPS

the term unity in the Equation [26]. It may be feasible to construct the locus for certain special cases where the transfer functions G_1, G_2, etc., are simple enough so that the relation may be combined into the following form without losing the identity of the original gain parameter (as outlined in the paper)

$$1 + G = 0 \quad [27]$$

which is the form from which the root-locus method of single-loop systems is developed. In general, the original gain parameter loses its identity when the characteristic equation is combined in the form of Equation [27].

The method indicated by the author for the complicated systems is similar to solving the problem loop by loop. By means of loop-by-loop approach it first determines the gain of the loop consisting of the blocks with the transfer functions G_3 and H_{33} from its equivalent transfer function $G_3/(1 + G_3H_{33})$ which is then combined with the blocks with the transfer functions G_2 and H_{32}. The gain of this second loop is then determined from the second equivalent transfer function. It is further proceeded to the third loop. With this approach there may not exist sufficient justification for the determination of the gain of the first loop before the subsequent determination of those second and third loops in the light of the performance of the over-all system.

Additional difficulties arise for the general application to multiloop systems. There is only one gain parameter being associated with the root locus, while there are usually several gain parameters that may be simultaneously selected in the multiloop system. Furthermore, the locus has to be replotted in case there are any changes in the time parameters in either single-loop or multiloop systems.

Although the original graphical method for constructing the locus for the case of multiloop systems is, in general, impracticable, yet the concept of presenting all possible roots of the characteristic equation on the complex plane is still plausible. It seems then that simple methods may be developed to compute the roots by means of existing calculating or computing machines and then to plot them on the complex plane. Another method has been suggested[3] which is to plot a "constrained root locus." This is developed by making use of the situation that only favorable sets of the roots are plotted from the combinations of adjustable gain and time parameters, without the necessity of the previous determinations of any gain of inner or first loops.

In conclusion, there may still exist a desirability to further advance the root-locus method for multiloop feedback-control systems.

R. H. Macmillan.[8] The author is to be congratulated upon the brevity, conciseness, and clarity which always mark his contributions to control theory. His name has been associated with so many fundamental advances in the theoretical field that it is becoming something of a byword and we are indeed fortunate in having him contribute to the Frequency Response Symposium. Perhaps the use of the root-locus method can be extracted from the realms of art and reduced more to a science by the use of the "phase-angle loci" introduced by Chu.[7,8] It is not denied that their use is more cumbersome than the intuitive application of experience, where this is possible, but it does make the root-locus approach more readily available to use by less skilled persons.

With regard to Mr. Evans' remarks concerning a desirable configuration for the roots, it is interesting to note that Dr. Whiteley in a paper published in 1946[9] found a satisfactory response was obtainable if the poles were distributed in arithmetical progression along the negative real axis.

It must always be remembered that the root-locus approach is applicable only to linear systems, though it is possible[8] to treat systems involving a finite time lag or a standard form of distributed time lag.

George Biernson.[10] The author's development of the root-locus method has been a great advance in feedback-control theory. A knowledge of the closed-loop poles enables one to understand certain effects of the system transient response which are not evident from the frequency response. It is helpful in explaining such effects as tails and multiple oscillations in the transients. Particularly, the writer has found that the behavior of certain complex multiloop control systems is much more understandable if the closed-loop poles are investigated.

On the other hand, the frequency-response approach still remains a very effective synthesis tool. The calculation of the closed-loop poles should supplement the frequency-response approach rather than replace it. This is not only because the frequency-response approach is much easier to apply, but also because the two major characteristics of the transient response, build-up time and peak overshoot, can generally be estimated better from the frequency response than from the values of the closed-loop poles.

To illustrate this conclusion, consider the load-damped servo which the author discusses. In practice, the damper is generally adjusted to have an attenuation ratio of roughly 9 (or slightly less), which corresponds to an 8-to-1 inertia ratio between the floating slug and the combined motor and load. To make the numbers easy, assume that the time constant of the damper is 1 sec. The complete open-loop transfer function, designated $G(s)$, would be approximately as follows

$$\frac{\theta_o}{\theta_e} = G(s) = \frac{3(s+1)9}{s^3(s+9)}$$

The particular gain setting shown yields an M_p (peak magnitude of $\theta_o(j\omega)/\theta_i(j\omega)$ of 1.30. For the best response one would probably raise the gain by about 50 per cent to yield the minimum M_p, but the setting shown is still quite reasonable. The closed-loop transfer function is

$$\frac{\theta_o}{\theta_e} = \frac{G}{1+G} = \frac{3(s+1)9}{(s+3)^3}$$

The closed loop thus has a triple-order real pole. Since there is no quadratic pole, one might expect that the system has no overshoot; but this is not so. In reference[11] the step response of this system is plotted in Fig. 3 of the discussion. Fig. 3 compares the response of this system to that of a second-order system having the same M_p (1.30) and the same band width, as defined by the frequency where the phase lag of θ_o/θ_i is 90 deg. Both transients have essentially the same amount of overshoot (25 per cent) and the same build-up time (the initial time for the transient to rise to unity).

The principle of the root locus can be applied in a much more convenient manner than that used by the author, by employing generalized frequency-response plots of log magnitude versus log frequency, as described in the reference. It is particularly convenient to recognize that the low-frequency closed-loop poles are approximately equal to the low-frequency open-loop zeros of $G(s)$. Consequently, it is the low-frequency zeros of $G(s)$ which cause tails in the transient response.

R. Oldenburger.[12] The author's well-known and excellent approach to control problems, discussed on the first two pages of the paper, concentrates on the roots of the characteristic equations. This is well founded since such equations exist for a very wide variety of control problems, and for these problems the nature of the transients can be seen readily from a glance at the roots of these equations (also frequency response follows readily from such roots and those of $\Theta_o/\Theta_i = 0$). Quick design thus depends on the rapid solution for these roots. The writer has tried all of the methods that he could locate in the literature. Synthetic division is by far the most rapid he found. His experience as an algebraist is that it simply takes too long to draw curves. Apparently, by the time an engineer gets to automatic-control problems he has forgotten synthetic division. Otherwise, more engineers would certainly use this approach. The writer will illustrate his technique for solving algebraic equations.

Consider the algebraic equation

$$x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n = 0 \ldots [28]$$

Write the equation

$$x + a_1 = 0$$

formed from the first two terms. This gives the trial root $x = -a_1$. We use the method of synthetic division introduced in the writer's June-July *American Mathematical Monthly* paper, referred to in the Frequency Response Symposium paper No. 53—A-11. This goes as follows

1	a_1	a_2	$\ldots$	a_{n-2}	a_{n-1}	a_n	$\lvert a_1$
.	.	.	.	$\left[a_{n-2} - \frac{1}{a_1}\left(a_{n-1} - \frac{a_n}{a_1}\right)\right]$	$\left(a_{n-1} - \frac{a_n}{a_1}\right)$	a_n	
R				$\frac{1}{a_1}\left(a_{n-1} - \frac{a_n}{a_1}\right)$	$\frac{a_n}{a_1}$	0	

Here we divide from right to left. This is the opposite of the left-to-right procedure in ordinary synthetic division. We as-

[8] University Lecturer, University of Cambridge, Cambridge, England.

[7] "Study of Process Control by Phase-Angle Loci Method," by Y. Chu, Instrument Society of America, 1953, not yet published.

[8] "Feedback Control Systems With Dead Time or Distributed Lag," by Y. Chu, Trans. AIEE, *Applications and Industry*, vol. 71, 1952, pp. 291–296.

[9] "Theory of Servo Systems," by A. L. Whiteley, *Journal of the Institution of Electrical Engineers*, London, England, vol. 93, part 2, 1946, pp. 353–367.

[10] Research Engineer, Servomechanisms Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

[11] "Quick Methods for Evaluating the Closed-Loop Poles of Feedback Control Systems," by G. Biernson Trans. AIEE, *Applications and Industry*, vol. 72, 1953, pp. 53–70.

[12] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

sume that $-a_1$ is a root and that the final remainder is 0 as shown. In practice we do not write this zero. This zero is the result of subtracting the entry a_n above it from the a_n at the head of the last column. In practice we do not write the a_n above the zero. We divide the a_n above the zero by the divisor a_1 to obtain the next to the last remainder a_n/a_1 shown. We subtract this remainder from the next to the last coefficient a_{n-1} to obtain

$$\left(a_{n-1} - \frac{a_n}{a_1}\right)$$

as shown. We divide this entry

$$\left(a_{n-1} - \frac{a_n}{a_1}\right)$$

by a_1 to obtain the third from the last remainder

$$\frac{\left(a_{n-1} - \dfrac{a_n}{a_1}\right)}{a_1}$$

Continuing this process, we finally arrive at the remainder R. If the division were exact this remainder plus the divisor a_1 would equal the entry a_1 in the second column. Naturally, it does not, and we try the divisor $a_1 - R$ instead of a_1, repeating the division process; that is, we assume that there is a root $R - a_1$. If division by $a_1 - R$ yields a remainder R_1 (in place of R), we can take $a_1 - R - R_1$ as the next divisor, etc. This method quickly yields the numerically large real roots, because after we have determined a root we use the remainders corresponding to an equation of one lower order and repeat the process. The first remainder is 1 and is not normally written.

To illustrate, consider the equation

$$x^3 + 121x^2 + 2030x + 1010 = 0$$

From the first two coefficients we get —121 as the trial root. We divide as follows

1	121	2030	1010	|121
		2022		
	17	8		

The number 8 is obtained by dividing 1010 by 121. Subtracting 8 from 2030 we obtain the entry 2022. Dividing 2022 by 121 we secure the entry 17. If the division were exact the sum 138 of the divisor 121 and the remainder 17 would be the entry 121 in the second column. Clearly, the divisor 121 is too large by 17. Therefore, for the next trial try 121 — 17, i.e., 104. We now have (rounding off some numbers)

1	121	2030	1010	|104
		2020		
	19	10		

Since 19 + 104, i.e., 123, is too large by the number 2, we try 104 — 2, i.e., 102 next. We then have

1	121	2030	1010	|102
		2020		
	20	10		

However, 20 + 102 is 122, which is 1 too large. We therefore try 102 — 1, i.e., 101, and obtain

1	121	2030	1010	|101
		2020		
	20	10		

Since the remainder 20 plus the divisor 101 equal the entry 121, the quantity —101 is a root of the given equation. From the remainders we form the equation

$$x^2 + 20x + 10 = 0$$

which must now be solved.

We remark that in practice the foregoing synthetic division would be reduced considerably. Thus, since there is more than one root, one knows that 121 is too large as a first guess and 100 is the obvious first try. The first try should be an easy number to work with and should be well rounded off. Then we have

1	121	2030	1010	|100
		2020		
	20	10		

Since 20 + 100, i.e., 120, is short by 1 we try 100 + 1, i.e., 101 next and find that this is the solution. This is the way the writer solved this problem when it came up in the study of a locomotive governor.

From the first two terms of the quadratic equation given we have the tentative root —20, i.e., the tentative divisor 20. We divide as shown

1	20	10	|20
	0.5		

Here 10 divided by 20 gives the remainder 0.5. Now the remainder 0.5 plus the divisor 20 is too large by 0.5, which is not much. In practice the writer would stop here and use —20 and —0.5 as the remaining roots, or more accurately —19.5 and —0.5.

By removing the numerically large real roots the problem generally can be cut down rapidly to a reasonable equation of low degree with smaller coefficients. That such roots generally exist follows from probability theory.

Numerically small real roots are found as follows: From the last two terms of the given equation we form the equation

$$2030x + 1010 = 0.$$

Solving this gives the trial root —0.5. Ordinary left-to-right synthetic division yields

1	121	2030	1010	|0.5
	0.5	60	985	
1	121	1970	25	

Since the remainder 25 is small we know that —0.5 is a root. In practice we would not solve for the numerically small real roots until the numerically large real roots had been removed.

After all real roots have been obtained we solve for the complex roots as described in the writer's *American Mathematical Monthly* paper on the subject. The division is done with pairs of numbers. It is more difficult than solving for real roots but can be done very rapidly by a skillful operator. Of course, we ordinarily use only one or two digits because any more would be absurd in view of the errors involved in the fitting of the mathematics to the physical system.

Solving the equation formed by the first three coefficients, i.e.

$$x^2 + a_1x + a_2 = 0$$

will often yield the numerically largest real root more accurately. It also may yield complex roots approximately.

Consider the equation

$$x^6 + 228x^5 + 6000x^4 + 37000x^3 + 210000x^2 + 320000x + 274000 = 0$$

Solution of

$$x^2 + 228x + 6000 = 0$$

gives the tentative root —198. Right-to-left division yields

1	228	6000	37000	210000	320000	274000	|198
		5800	36000	208000	319000		
	29	182	1050	1610	1380		

The sum of 29 and 198 is 227. We should have 228. This suggests the divisor 199. We have

1	228	6000	37000	210000	320000	274000	|199
		5800	36000	208000	319000		
	29	181	1050	1600	1380		

Thus —199 is a root. The coefficients of reduced equation follow with the division by 23 shown

1	29	181	1050	1600	1380	|23
		140	980	1500		
	6	43	65	60		

Clearly —23 is a root. The next reduced equation is

$$x^4 + 6x^3 + 43x^2 + 65x + 60 = 0 \quad \ldots\ldots [29]$$

The first three terms of Equation [29] yield the trial factor

$$x^2 + 6x + 43$$

We divide from right to left as shown

1	6	43	65	60	1, 6, 43	
		1.4	8		1, 1.3, 1.4	
		42	57			
	1.3	8				
	4.7	34				
	6	43				
	—1.3	—9				

The entries are obtained as follows: Assuming that the final remainder is zero, we divide the 43 of the divisor into the last entry 60 of the dividend to yield the "quotient" 1.4. We multiply 1.4 by the 1 and 6 of the divisor to yield the entries 1.4 and 8 written beneath 43 and 65, respectively. We subtract and obtain the 42 and 57 in the third row. Division of 57 by 43 yields the quotient 1.3. We multiply 1.3 by the 1 and 6 of the divisor to obtain the entries 1.3 and 8 in the fourth row. Subtracting these entries from 6 and 42 we have the remainders 4.7 and 34, respectively. The next division is the last, and the quotient should be 1. We therefore write 6 and 43 beneath the 4.7 and 34 as shown. Subtracting, we get the remainders —1.3 and —9. If the division were exact these remainders would be zeros. We have subtracted off too much. We therefore decrease the 6 and 43 by 1.3 and 9, respectively, and obtain $x^2 + 4.7x + 34$ as the next trial divisor. We have the division

1	6	43	65	60	1, 4.7, 34	
		1.8	8		1, 1.7, 1.8	
		41	57			
	1.7	8				
	4.3	33				
	4.7	34				
	—0.4	—1				

Since the remainders are numerically small, the factor

$$x^2 + 4.7x + 34$$

of the left side of Equation [29] is very nearly correct. The quotient obtained is $x^2 + 1.7x + 1.8$ so that the left member of Equation [29] factors approximately into

$$(x^2 + 4.7x + 34)(x^2 + 1.7x + 1.8)$$

Further division yields the correct factorization

$$(x^2 + 4.3x + 32.5)(x^2 + 1.7x + 1.8)$$

but the previous factorization is good enough for ordinary practical purposes.

For his initial division in this problem an intelligent analyst would try coefficients less than 6 and 43, such as 5 and 30.

As in Lin's method[12] the factor $(x^2 + 1.7x + 1.8)$ could have been obtained from the last three terms of the polynomial in Equation [29], and left-to-right division. From the last three terms we can factor 43 to obtain the trial divisor $(x^2 + 1.5x + 1.4)$. The division is (see the writer's *American Mathematical Monthly* paper)

1.5, 1.4	1	6	43	65	60
		1.5	1.4		
		4.5	42		
			7	6	
		35	59		
			52	49	
			7	11	

The remainders are both positive. This shows that we did not subtract enough, and should increase the "divisors" 1.5 and 1.4.

We summarize the writer's synthetic division method of solving an algebraic Equation [28]. It is assumed that the equation is associated with an automatic control problem. If the equation is of high degree we first apply Routh's criterion to insure that the equation represents a stable system (roots have negative real parts).

We write Equation [28] as

$$L_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_{n-2} x^2 + a_{n-1} L_2 = 0 \ldots [30]$$

and as

$$Q_1 x^{n-2} + a_3 x^{n-3} + \ldots + a_{n-2} x^2 + a_{n-2} Q_2 = 0 \ldots [31]$$

where

$$L_1 = x + a_1$$

$$L_2 = x + \frac{a_n}{a_{n-1}}$$

$$Q_1 = x^2 + a_1 x + a_2$$

$$Q_2 = x^2 + \frac{a_{n-1}}{a_{n-2}} x + \frac{a_n}{a_{n-2}}$$

1 Use the linear expression L_1 as a trial factor, or better, $x + \alpha$ where α is a conveniently chosen number less than a_1. Employ right-to-left synthetic division. Correct α on the basis of the remainder. When a root has been found, work with the reduced equation (one lower degree). In this way solve for the numerically large real roots and obtain a reduced equation E_2.

2 Write the reduced equation in the form of Equation [30]. Use the linear expression L_2 for the reduced equation E_2 as a trial factor, or better, $x + \beta$ where β is greater than a_n/a_{n-1}. Use ordinary left-to-right synthetic division. In this way solve for the rest of the real roots (the numerically small real roots).

3 When all of the factors corresponding to real roots are removed, write the reduced equation as in Equation [31]. With Q_1

[12] "Instrument Engineering," by C. S. Draper, W. McKay, and S. Lees, McGraw-Hill Book Company, Inc., New York, N. Y., 1953, p. 427.

as a trial divisor, or better, $x^2 + \gamma x + \delta$ where γ and δ are less than a_1 and a_2, respectively, and right-to-left synthetic division, remove a quadratic factor, correcting trial divisors on the basis of remainders. Quadratic factors with relatively "large" coefficients are readily found by this procedure.

4 If a quadratic factor is not easily found by step 3 suppose that Equation [31] is the reduced equation obtained after the foregoing steps. With Q_2 as a trial divisor, or better, $x^2 + \lambda x + \mu$ where λ and μ are greater than a_{n-1}/a_{n-2} and a_n/a_{n-2}, respectively, and left-to-right synthetic division, remove a quadratic factor, correcting trial divisors on the basis of remainders. Quadratic factors with relatively "small" coefficients are readily found by left-to-right synthetic division.

Thus right-to-left synthetic division is to be used to remove linear and quadratic factors corresponding to L_1 and Q_1 obtained from the high-order terms of the equation, and left-to-right synthetic division for removing linear and quadratic factors L_2 and Q_2 corresponding to low-order terms of the equation.

A big advantage of right-to-left synthetic division is that the correction to be made in a trial divisor is immediately evident. At the end of a division one knows how far one is away from the actual roots. There is probably no one best method for solving all algebraic equations. It is the writer's experience that intelligent cut and try is better than one definite procedure, such as that of Lin for finding complex roots, wherein each step is prescribed and one can without the use of judgment proceed mechanically from start to conclusion.

One cannot judge the value of a method, such as that of the author's or the writer's, on the basis of one problem. To judge a method one must work several problems, as they occur in industry, so that one becomes adept at using the method and acquires the "feel" of it. Only in this way can the efficacy of one method be compared correctly with another.

Author's Closure

The author appreciates the discussions because they point out the broader problem and the appropriate methods of attack. In particular, it is worth while to note the manner in which simplifications of the problem itself are exploited in the various methods. For the convenience of the reader, the comments will follow the pattern of the individual discussions.

Yaohan Chu's discussion gives root-locus undue credit in that item (*a*), the plotting of all roots on a complex plane, is given in Bode's book.[14] The characteristic equation $KG(s) = -1$ arises naturally for a single loop system in which the zeros and poles of $G(s)$ are known and K is the variable of interest. The frequency-response method first exploited knowing the factors of $G(s)$ by treating each factor as a vector so that the product $G(s)$ has a magnitude equal to the product of the magnitudes and a phase angle equal to the sum of the phase angles of the individual factors. Root-locus uses this same simplification for complex values of s by plotting vectors in the s-plane.

The organization of the solution of a complex system in the simplest form benefits either a frequency-response or a root-locus solution. If the problem is as stated in Equation [18], for example, then the adjustable functions $G_{31}(s)$ and $G_{32}(s)$ are isolated in the expansion given in Equation [19]. The effect of the functions can then be studied in individual root-locus plots or the frequency-response curves could be added separately to the curve of $F_{33}(s)\,M_{33}(s)$. If, however, the problem is as given in Equation [26] with G_1 and G_2 adjustable, then the term involving the product G_1G_2 complicates either method of solution. Fortunately, Equation [18] arises often in that the pitch system of an airplane, for example, involves moment, lift, and control equations with only the control elements adjustable. A roll-yaw system can be solved as roll and yaw separately with the effects of intercoupling added. These intercoupling calculations are benefited by Chu's suggestion that only critical roots be considered. Analog computers are used widely for getting many such numerical results quickly, but a need still exists for a way of picturing a system which will permit organizing the analog study and provide a rough check on the results. This problem is sufficiently different from the original single-loop problem so that the best solution may not involve root-locus at all.

Mr. Macmillan's kind remarks about several fundamental advances suggest that a reference be given to the paper which expounds Dr. Profos' idea of conformal mapping from a Nyquist curve to obtain the dominant roots.[15] This procedure, in the author's opinion, still seems to be the logical one for a person who thinks in terms of frequency response but wants a rough check on transient response. The references to Chu's papers are welcome, particularly since an extension is given to the dead-time problem[3] not covered elsewhere in the literature. The limitation of root-locus to linear systems certainly should be emphasized because linearity is essential to describing the system by an algebraic equation in s.

George Biernson's example points out several reasons for caution in using zeros and poles. The zeros and poles completely define either frequency or transient response, but perhaps the emphasis on poles alone could cause many engineers to expect no transient overshoot without the complex poles. The closed-loop transfer function given is, however, equivalent to a critically damped system $1/(1 + s/3)^2$ followed by a lead network $(1 + s)/(1 + s/3)$. The lead-network function can be expressed as $1 + 2[1 - 1/(1 + Ts)]$ which shows that the actual transient rises above that of the critically damped system by twice the effect of an additional lag $1/(1 + Ts)$. Another caution is that roots, or closed-loop poles, shift very rapidly for a small change in a parameter when the roots are nearly equal. Specifically, the rate of change of s with respect to gain is infinite for two equal roots, and even the second derivative of s is infinite for three equal roots. A six per cent change in gain in the given system shifts a pair of roots to an undamped natural frequency of 4.3 with a damping ratio of 0.8 and shifts a real root to —1.5. The transient can now be plotted as a simple quadratic with the effect of the equivalent lead network cut roughly to $^1/_4$. The rapid shift of roots when they are nearly equal makes their numerical value hard to find, but the value is also not critical for either frequency response or transient response.

R. Oldenburger's synthetic division method deals with the classic problem of finding the roots of an nth degree polynomial equation. It is noted in Equation [19] that the minors of the determinant have all been expanded to polynomials. The advantage is even greater for a fourth order in which, in general, twenty-four terms would arise in the expansion. Substitution of numerical values permits cutting down the number of plots to the number of functions of interest. From having spent more time than the author cares to realize on graphical methods, however, he cannot help but accept the challenge of presenting an alternative to an algebraic solution. Thus in the first example the magnitude of each term

[14] "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand Company, Inc., New York, N. Y., 1945, p. 119.

[15] "Graphical Analysis of Control Systems," by W. R. Evans, Trans. AIEE, vol. 67, 1948, pp. 547–551.

can be plotted vertically on log-log co-ordinates versus the magnitude of x horizontally by straight lines, as shown in Fig. 9, herewith.

It is noted that at $|x| = 0.5$ the x^2 and x^3 terms are both small so that the characteristic function is nearly zero by making $x = -0.5$ so that the constant and the x-term cancel. Similar intersections give $x = -17$ and $x = -120$ as the first approximation to the other roots. In plotting these lines it was convenient to use regular squared paper with a logarithmic scale marked on a separate card. This card can then be used at $|x| = 17$ to note that the x^3 term is $^1/_8$ of the x and x^2 terms at this point. The root should be increased by $^1/_6$ to make the x^3 term be equal to the sum of the x and x^3 terms, making the root approximately -19.

Complex roots would arise if the x^2 term had been $12x^2$ rather than $120x^2$ which leaves the x and x^3 terms the largest and equal at $|x| = 45$. As a first approximation x would equal $j45$ to make these two terms cancel. Treating the magnitude of x and x^3 as unity, the logarithmic scale shows the $12x^2$ term to be 0.25. A separate vector diagram will show that x should have a damping ratio of 0.25/2 in order that the vector sum of the three terms be zero. A further approximation would involve noting that an increase in ζ would make the x and x^3 terms overcompensate for the x^2 term by twice $\sqrt{1-\zeta^2}$ times dx. The derivative of the quadratic function thus obtained can be divided into the remaining constant term of 1000 to get the additional increment in s needed to cancel the constant term.

R. Oldenburger's conclusions are certainly shared that one cannot judge a method on the basis of one problem and that one must become adept at all of them to compare them. In spite of our difference in preference between algebraic and graphical solutions, the author shares his preference on a more basic point that some judgment in a trial-and-error solution saves a lot of routine calculation.

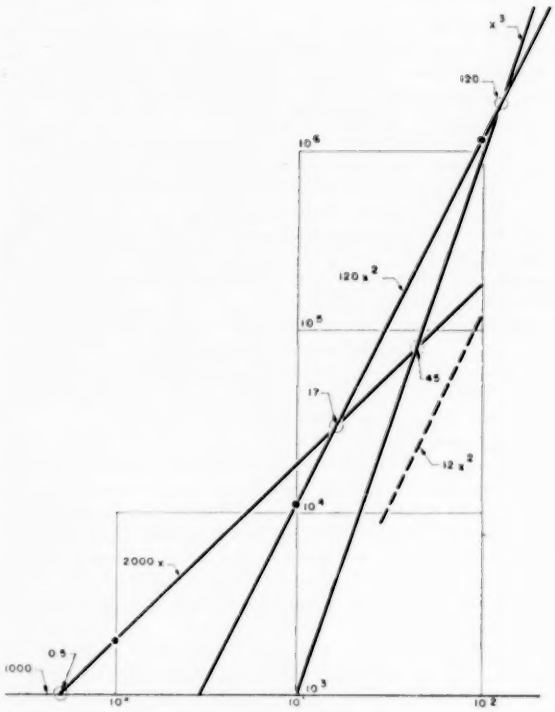


FIG. 9 GRAPHICAL SOLUTION OF $x^3 + 120x^2 + 2000x + 1000 = 0$

Approximate Frequency-Response Methods for Representing Saturation and Dead Band

By HAROLD CHESTNUT,[1] SCHENECTADY, N. Y.

This paper indicates there are numerous nonlinear elements having saturation or dead-band characteristics for which an "equivalent linear coefficient" may be used as the describing function. The equivalent gain factors for saturation and dead band are shown to have a somewhat complementary nature and to be of relatively simple form. Although the results of using this approximate method of analyzing backlash and saturation do not provide an exact solution, they do permit the designer to obtain useful qualitative results. As such, this method is particularly helpful for use in the more complex systems such as are generally encountered in practice, but that do not lend themselves to more exact analytical treatment.

Introduction

THE works of Kochenburger (1)[2] and Johnson (2) on applying frequency-response methods to nonlinear control systems can be applied in an approximate form to the problems of saturation and dead band. Each of these forms of nonlinearity can be shown to be capable of representation by an "equivalent linear coefficient" for its describing function. Typical examples of saturation and dead band are used to illustrate this physical approach to the problem.

The equivalent gain change for these nonlinear effects is evaluated for sinusoidal inputs. A complementary relationship between the saturation and dead-band gain-factor changes is indicated. The performance obtained while synchronizing to a step-position input for a servomechanism system having either saturation or dead band present will be compared with the results predicted by the use of the equivalent linear-coefficient form of describing function. Both exact and approximate analysis methods show that sustained oscillations can be maintained under certain conditions when saturation or backlash is present.

In addition to providing qualitative information on system performance with saturation or backlash present, the method outlined in this paper can be used to provide approximate results in handling nonlinear effects in more complicated, multiple-loop systems.

Typical Saturation and Dead-Band Problems

To provide a physical basis for the treatment of saturation and dead band that follows, a description of a few representative problems of each will be presented.

Saturation. The linear relationship between the output and the input of an amplifying device whether it be mechanical, electrical, or hydraulic is at best valid only over a limited range of values of the input function. For inputs greater than some limiting value, the output fails to continue to increase proportionally to the input and the effective output/input ratio is reduced. This phenomenon is known as saturation. The solid curve in Fig. 1 shows a form which the saturation characteristic may take. The input and the output may be in any appropriate units provided the singular valued relationship between output and input exists. To illustrate the effect of such a saturation characteristic on the describing function of an element, the case of voltage saturation in an electronic amplifier and of torque saturation in a d-c motor will be described.

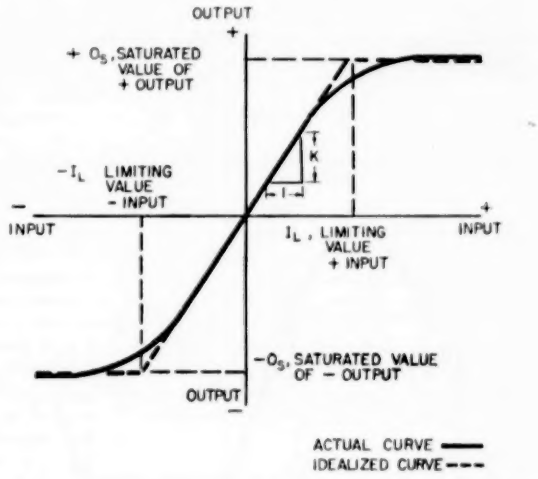


Fig. 1 Saturation Characteristic for Representative Control Element

Voltage Saturation in Electronic Amplifier. To provide a means for simplifying the analysis of the saturation problem in an electronic amplifier, it is assumed that the actual input-output curve is symmetrical and can be approximated over the region of operating interest by an "idealized" saturation characteristic, as shown in Fig. 1, by the dotted intersecting straight lines. One line passes through the origin and has the same slope, K, as does the actual gain characteristic in this region. The other two lines have zero slope and their values, $+O_s$ and $-O_s$, are approximately equal to the saturated value of the output, respectively, for large positive and large negative values of the input. I_L, the limiting value of the input is that input value where the lines of slope K and of zero slope intersect and beyond which the output can be considered no longer to increase. Thus

$$I_L = \frac{O_s}{K} \quad \ldots\ldots\ldots\ldots\ldots\ldots [1]$$

Symmetry has been assumed so that the limiting input for positive and negative inputs is the same and the corresponding positive and negative saturated values of output are equal in magnitude.

[1] Engineer, General Electric Company.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 19, 1953. Paper No. 53—A-25.

If, to an amplifier having the idealized input-output characteristic of Fig. 1, a sinusoidal input

$$i_a(t) = I_M \sin t \quad \ldots\ldots [2]$$

is impressed, in which $I_M > I_L$, the output $O_a(t)$ will have the form shown in Fig. 2.

For values of $-I_L < i_a < +I_L$, the output $O_a(t)$ will be proportional to $i_a(t)$ the proportionality factor being K. For values of $i_a > I_L$, the output will be constant and equal to O_s the saturated value of O_a; likewise, for values of $i_a < -I_L$, the output will be constant and equal to $-O_s$. The angle α at which $O_a(t)$ is no longer linearly proportional to $i_a(t)$ is

$$\alpha = \sin^{-1} \frac{I_L}{I_M} \quad \ldots\ldots [3]$$

It should be noted that for an electronic amplifier the form of

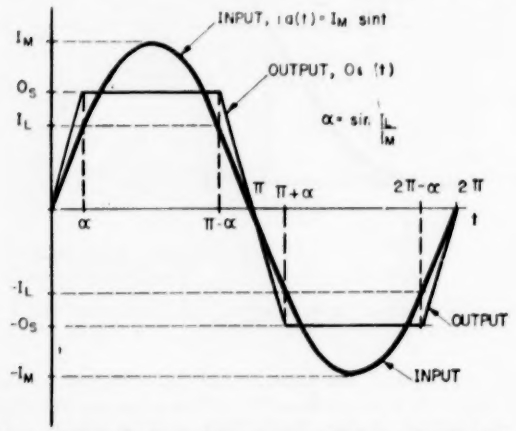


FIG. 2 OUTPUT $O_a(t)$ AS FUNCTION OF TIME FOR INPUT $i_a(t) = I_M \sin t$ TO A SATURATED ELEMENT WITH $I_M > I_L$, THE LIMITING VALUE OF INPUT

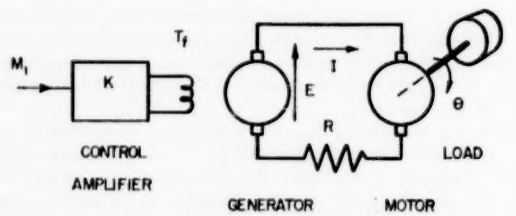


FIG. 3 MOTOR-GENERATOR CONTROL HAVING MOTOR-TORQUE SATURATION

the curves in Fig. 2 and the value of α are independent of the frequency of the input sinusoid for the range of frequencies that are of interest. For such sinusoidal inputs, the fundamental of the output, O_{a1}, is in phase with the fundamental of the input, I_M, and the ratio between them is a numeric that is a function of the magnitude of α. Thus

$$\frac{O_{a1}}{I_M} = G'(\alpha) \quad \ldots\ldots [4]$$

where $G'(\alpha)$ is the describing function of the saturated amplifier. Since α and I_L/I_M are related by Equation [3] it is apparent that O_{a1}/I_M of Equation [4] also could be expressed in terms of a function of I_L/I_M rather than of α.

Motor-Torque Saturation. The effect of motor-torque saturation

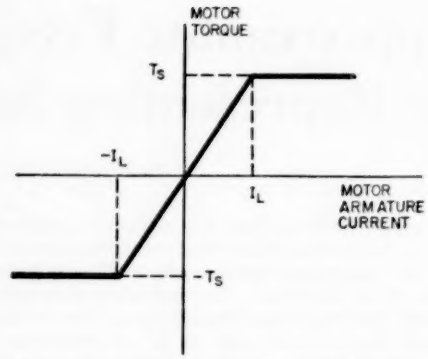


FIG. 4 IDEALIZED MOTOR TORQUE-ARMATURE CURRENT-SATURATION CHARACTERISTIC

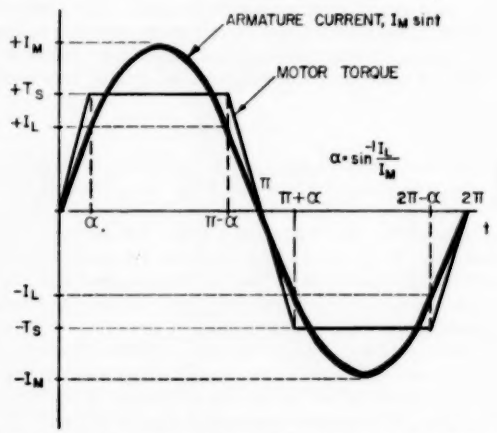


FIG. 5 MOTOR TORQUE AND ARMATURE CURRENT AS A FUNCTION OF TIME FOR SINUSOIDAL VARIATION IN ARMATURE CURRENT WITH TORQUE SATURATION PRESENT

on a d-c motor-generator control of the sort shown in Fig. 3, may produce sluggish and otherwise objectionable control-system performance. Owing to demagnetization effects caused by incomplete compensating windings and by magnetic saturation, the motor torque obtainable as a function of armature current may have the idealized saturation characteristic shown in Fig. 4, where T_s is the saturated value of torque available. Fig. 5 shows the armature current and corresponding motor torque as a function of time for a sinusoidal variation in armature current of so large a magnitude as to produce torque saturation. The fundamental of the torque is in phase with the armature current and is of a magnitude that is a function of I_M, the maximum of the armature current. Thus the fundamental motor torque can be expressed as

$$\text{torque} = K_T' I \quad \ldots\ldots [5]$$

where K_T' is the effective torque constant and has a value of K_T for $-I_L < I_M < +I_L$ and $K_T > K_T' > 0$ for $I_M > I_L$ and $I_M < -I_L$.

Referring to Fig. 3, the control, voltage, and torque equations can be expressed as

$$K M_1 = (T_f s + 1) E \quad \ldots\ldots [6]$$

$$E = RI + K_e s \theta \quad \ldots\ldots [7]$$

$$K_T' I = J s^2 \theta \quad \ldots\ldots [8]$$

where

M_1 = volts input to control amplifier
K = control-amplifier output volts per unit of input
T_f = time constant of control amplifier

From Equations [6] to [8]

$$\frac{\theta}{M_1} = \frac{K/K_e}{s(T_f s + 1)(T_M' s + 1)} \quad \ldots\ldots\ldots [9]$$

where

$$T_M' = \frac{JR}{K_T' K_e}$$

the effective motor time constant.

From this expression for the motor time constant it appears that T_M' increases with increasing amounts of torque saturation; i.e., lower values of K_T'. The undesirable effects that may result from increasing the time constant of a motor are well known and may prove objectionable from a control point of view.

Dead Band. Dead band or backlash is another form of non-linearity that is commonplace in control systems. Although dead band is generally present to some degree because of the limitations of the manufacturing process, there also are instances where an effect similar to dead band is introduced for purposes of providing limiting action. The characteristics of dead band are illustrated graphically in Fig. 6, where the dotted line shows an idealized plot of the output of an element with dead band as a function of the element input. The solid line shows more nearly an actual output-input characteristic in which the output/input slope changes gradually about $\pm I_B/2$ rather than changing abruptly at the point of discontinuity $\pm I_B/2$, as is indicated by the idealized "dotted" lines. For inputs $I > I_B/2$, the ratio of output to input is linear with the proportionality factor K being

$$K = \frac{\text{output}}{I - I_B/2} \quad \ldots\ldots\ldots [10]$$

If to an element having the idealized dead-band characteristic of Fig. 6, a sinusoidal input

$$i_a(t) = I_M \sin t \quad \ldots\ldots\ldots [11]$$

is impressed in which $I_M > I_B/2$, the output $O_a(t)$ will have the form shown in Fig. 7. For values of $-I_B/2 < i_a < +I_B/2$, i.e., in the dead-band region, the output $O_a(t)$ will be zero. For values of $i_a > \pm I_B/2$ the output will be proportional to $i_a - I_B/2$ with K from Equation [10] being the proportionality factor; for $i_a < -I_B/2$ the output is proportional to $i_a + I_B/2$ again with K as the factor for proportionality. The angle α at which $O_a(t)$ no longer is zero is

$$\alpha = \sin^{-1} \frac{I_B}{2I_M} \quad \ldots\ldots\ldots [12]$$

From Fig. 7, it will be noted that α is independent of the frequency of the input sinusoid and that the fundamental of the output is in phase with that of the input. The magnitude of the output is a function of the $I_B/2I_M$ ratio.

It should be noted that the input quantity refers to the input to the particular element and in some cases may represent a difference quantity or error rather than the entire input to the control system. This is the case in the instance of mechanical backlash as illustrated in the following example.

Spring-Load Backlash Including Drive Motor Effects. Fig. 8 shows schematically a d-c motor connected through backlash and a springy shaft to an inertia load having viscous damping present. The describing function for these elements can be developed with the aid of Figs. 9 and 10. As shown in Fig. 9, the difference in position between the motor θ_M and the load θ_L determines the torque exerted by the spring. For values of $-\theta_B/2 < (\theta_M - \theta_L) < +\theta_B/2$ there is no spring torque transmitted from motor to load. If the angle $\theta_M - \theta_L$ is considered to be sinusoidal in

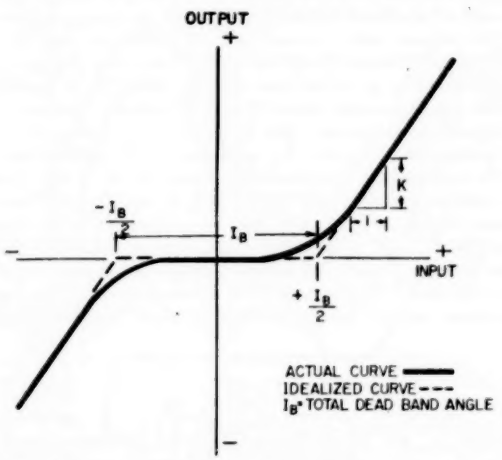


FIG. 6 OUTPUT VERSUS INPUT CHARACTERISTIC FOR AN ELEMENT WITH DEAD BAND OR BACKLASH

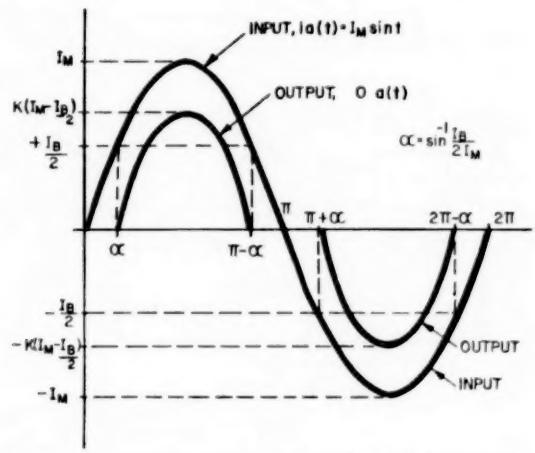


FIG. 7 OUTPUT $O_a(t)$ AS A FUNCTION OF TIME FOR INPUT $i_a(t) = I_M \sin t$ TO AN ELEMENT WITH DEAD BAND, SHOWN FOR $I_M > I_B/2$

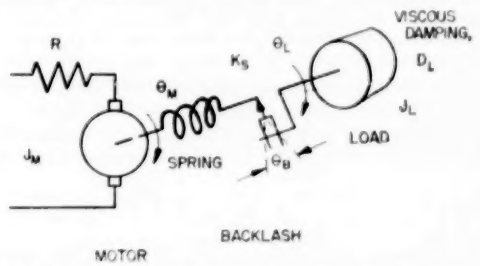


FIG. 8 DIRECT-CURRENT MOTOR CONNECTED THROUGH BACKLASH AND A SPRINGY SHAFT TO AN INERTIA LOAD HAVING VISCOUS DAMPING

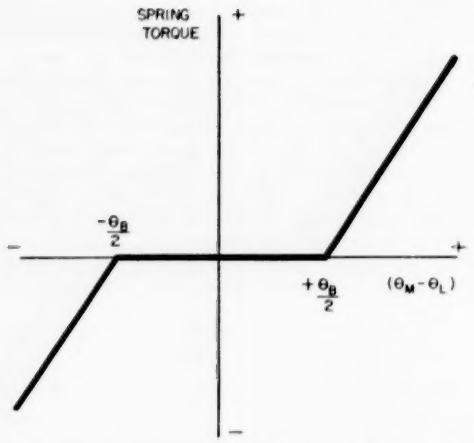


FIG. 9 TORQUE DEVELOPED ACROSS SPRING IN FIG. 8, AS A FUNCTION OF DIFFERENCE IN MOTOR AND LOAD SHAFT POSITION WITH TOTAL BACKLASH θ_B

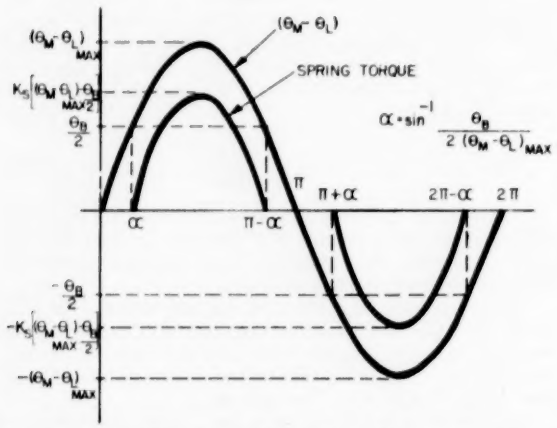


FIG. 10 SPRING TORQUE AND MOTOR LOAD-POSITION DIFFERENCE ANGLE AS A FUNCTION OF TIME FOR SINUSOIDAL VARIATION IN MOTOR-LOAD DIFFERENCE ANGLE WITH TOTAL BACKLASH, θ_B

nature, the spring torque transmitted will have the form shown in Fig. 10. From this it can be seen that the fundamental of the spring torque is in phase with the input angular difference $(\theta_M - \theta_L)$, the proportionality factor K_s' being a function of the ratio of

$$\frac{\theta_B}{2(\theta_M - \theta_L)_{max}}$$

In essence K_s' is an effective linear spring constant, the value of which is a function of the amplitude ratio noted.

Using an equivalent linear spring coefficient K_s', the equation for the torques acting on the load is

$$J_L s^2 \theta_L + D_L s \theta_L = K_s' (\theta_M - \theta_L) \quad \ldots\ldots [13]$$

The torques acting on the motor are

$$K_T I - K_s' (\theta_M - \theta_L) = J_M s^2 \theta_M \quad \ldots\ldots [14]$$

and the motor-voltage equation is

$$E = RI + K_s s \theta_M \quad \ldots\ldots [15]$$

Combining Equations [14] and [15]

$$\frac{K_T E}{R} - K_s' (\theta_M - \theta_L) = J_M s^2 \theta_M + \frac{K_s K_T}{R} s \theta_M \quad \ldots [16]$$

Equations [13] and [16] are shown in block-diagram form in Fig. 11, where K_s' is shown by the schematic feedback symbol. For purposes of analytical work, K_s' may be considered to have an effective linear value that is dependent on the

$$\frac{\theta_B}{2(\theta_M - \theta_L)_{max}}$$

ratio.

Current Limiting Action. A somewhat different form of the dead-band problem is illustrated in Fig. 12, where is shown a motor-generator control in which a current-limit feedback is provided to prevent excessive motor current. The output of the limiter amplifier is supplied degeneratively to the control amplifier thereby limiting the generator voltage and consequently the motor current. This example of limiting action is similar in its method of treatment to that for velocity, acceleration, pressure, or other parameter limiting that may be employed with the appropriate quantity being fed back through the limit amplifier.

The voltage-output characteristic of the limiter amplifier as a function of the motor-armature current is shown in Fig. 13. For positive values of current less than I_L, the value of current at which limiting action starts, there is no voltage feedback. For values of current greater than I_L, the feedback voltage increases proportionally to current, K_H being the proportionality factor.

For a sinusoidal variation in motor-armature current greater in magnitude than I_L, the voltage from the limiter amplifier is as shown in Fig. 14. The fundamental component of the limiter-amplifier voltage is in phase with the motor-armature current. The magnitude of the fundamental component of limiter voltage is a function of the maximum of the magnitude of the motor-armature current, the proportionality factor being K_H', where

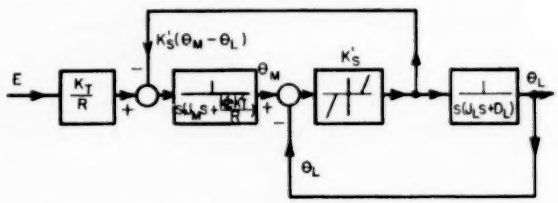


FIG. 11 BLOCK DIAGRAM FOR ELEMENTS SHOWN IN FIG. 8

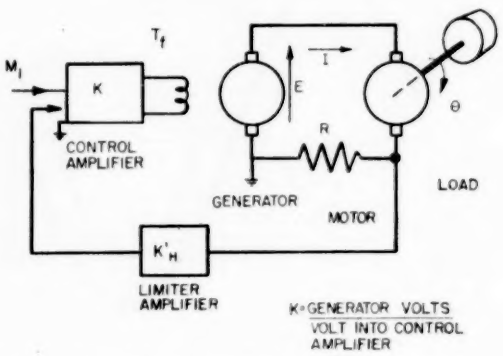


FIG. 12 MOTOR-GENERATOR CONTROL WITH CURRENT-LIMIT FEEDBACK

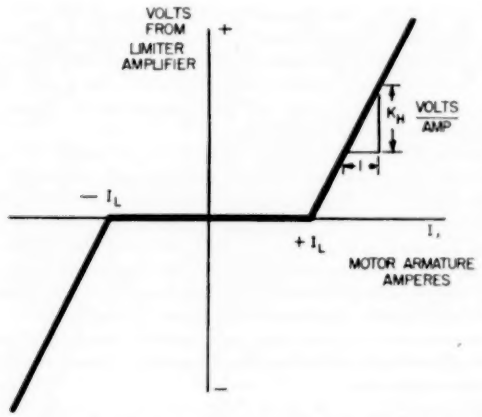


FIG. 13 LIMITER-AMPLIFIER VOLTAGE CHARACTERISTIC AS A FUNCTION OF MOTOR-ARMATURE CURRENT FOR CURRENT-LIMIT ACTION

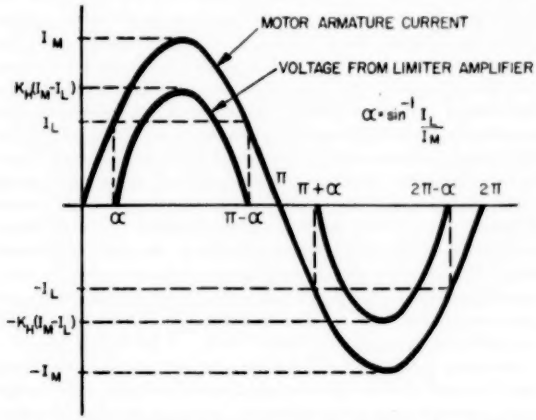


FIG. 14 MOTOR-ARMATURE CURRENT AND VOLTAGE FROM LIMITER AMPLIFIER AS A FUNCTION OF TIME FOR SINUSOIDAL VARIATION IN MOTOR-ARMATURE CURRENT; LIMIT ACTION TAKES PLACE AT I_L

$O < K_H' < K_H$. Thus the limiter-voltage output can be expressed as $K_H'I$.

Referring to Fig. 12, the following equations may be written:

$$\text{Signal to amplifier} = M_1 - K_H'I \quad [17]$$

$$\text{Generator voltage, } E = \frac{K(M_1 - K_H'I)}{(T_f s + 1)} = RI + K_e s\theta \quad [18]$$

$$\text{Motor torque, } K_TI = J s^2 \theta \quad [19]$$

Combining these equations, the transfer function from input-signal voltage to motor position is

$$\frac{\theta}{M_1} = \frac{K}{K_e s\left[T_f T_M s^2 + \left(T_M + T_f + \frac{KK_H'J}{K_eK_T}\right)s + 1\right]} \quad [20]$$

where

$$T_M = \frac{JR}{K_eK_T}$$

With $K_H' = 0$, this transfer function takes on the usual value it has with no limiting action. With limiting action present and $K_H' \cong K_H$, the transfer function has a more sluggish response caused by the

$$\frac{KK_H'Js}{K_eK_T}$$

term in the denominator.

EVALUATION OF GAIN-CHANGE FACTOR

The preceding examples of typical saturation and dead-band problems have served to point out two common types of nonlinear characteristics for which it is desirable to relate the magnitude of the fundamental of the output to the magnitude of a sinusoidal input. This ratio can be considered as a gain-change factor for it indicates the factor by which the gain of a linear element must be modified to account for the effect of the nonlinearity present. By multiplying the linear coefficient for the element by the gain-change factor, the value for the "equivalent linear coefficient" of the element having a given input amplitude and region of nonlinearity can be found. Thus if A is the gain-change factor, and K the linear coefficient for the element, the equivalent linear coefficient K' is

$$K' = AK \quad [21]$$

where A is a function of the input amplitude.

Saturation-Gain Factor, A_s. Although the saturation-gain factor has been developed by Kochenburger (3), it will be repeated here to help point the complementary nature of the saturation and dead-band gain factors. Fig. 2 shows the output of a symmetrically saturated element as a function of time for a sinusoidal input. From this it is evident that there is no phase shift between the fundamental of the output with respect to the input fundamental. With K as the gain of the element in the linear region, I_L the value of input at which limiting action takes place, and I_M the maximum value of the input, the output for a half cycle can be expressed as follows

$$\begin{aligned} O_a(t) &= K\,I_M \sin t & 0 < t < \alpha \\ O_a(t) &= K\,I_M \sin \alpha & \alpha < t < \pi - \alpha \\ O_a(t) &= K\,I_M \sin t & \pi - \alpha < t < \pi \end{aligned}$$

where

$$\alpha = \sin^{-1}\frac{I_L}{I_M} \quad [22]$$

Taking advantage of symmetry O_{a1}, the fundamental of the output can be expressed

$$O_{a1} = \frac{2K\,I_M}{\pi}\left(2\int_0^{\alpha} \sin^2 t\,dt + \int_{\alpha}^{\pi-\alpha} \sin\alpha \sin t\,dt\right)$$

$$O_{a1} = K\,I_M\left(\frac{2\alpha}{\pi} + \frac{\sin 2\alpha}{\pi}\right) \quad [23]$$

Defining the saturation gain factor A_s as the ratio of the fundamental of the element with saturation to the fundamental of the same element with linear operation throughout

$$A_s = \frac{2\alpha}{\pi} + \frac{\sin 2\alpha}{\pi} \quad [24]$$

Fig. 15 is a plot of A_s (left ordinate) as a function of I_L/I_M and shows the gain factor to fall off from 1.0 slowly for $I_L/I_M > 0.8$, while for small values of I_L/I_M (< 0.3) the gain factor is approximately proportional to I_L/I_M. As indicated by the title of the right ordinates in Fig. 15, and as will be shown in the material that follows, the saturation and dead-band gain factors are somewhat complementary in nature.

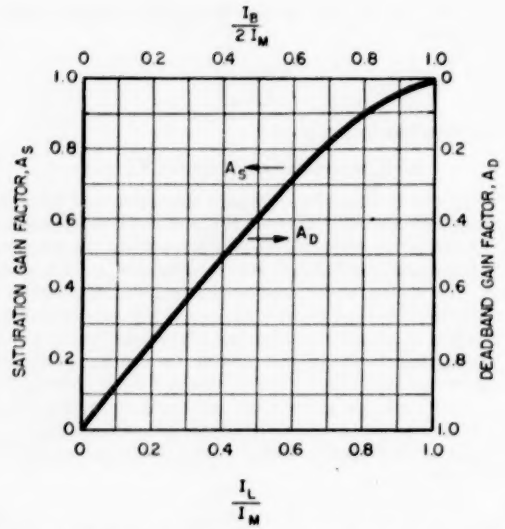


FIG. 15 SATURATION-GAIN FACTOR A_S, AND DEAD-BAND GAIN FACTOR A_D, AS A FUNCTION OF (INPUT LIMIT VALUE/MAXIMUM VALUE)

Dead-Band Gain Factor, A_D. Fig. 7 shows the output of an element with a symmetrical dead-band characteristic as a function of time for a sinusoidal input. Again the fundamental of the output is in phase with the fundamental of the input as was the case with saturation. With K as the gain of the element in the linear region, $I_B/2$ the value of input at which linear operation begins, and I_M the maximum value of the input, the output for half a cycle can be expressed as follows

$$O_s(t) = K I_M (\sin t - \sin \alpha) \quad \alpha < t < \pi - \alpha$$

where

$$\alpha = \sin^{-1} \frac{I_B}{2I_M} \ldots\ldots\ldots\ldots\ldots\ldots [25]$$

$O_s(t) = 0$ for the remaining interval; O_{s1}, the fundamental of the output, can be expressed as

$$O_{s1} = \frac{2K I_M}{\pi} \int_{\alpha}^{\pi-\alpha} (\sin t - \sin \alpha) \sin t \, dt$$

$$= KI_M \left(1 - \frac{2\alpha}{\pi} - \frac{\sin 2\alpha}{\pi}\right) \ldots\ldots [26]$$

Defining the dead-band gain factor as the ratio of the fundamental of the element with the dead band to the fundamental of the same element with linear operation throughout

$$A_D = 1 - \frac{2\alpha}{\pi} - \frac{\sin 2\alpha}{\pi} \ldots\ldots\ldots\ldots [27]$$

or

$$A_D = 1 - A_s \ldots\ldots\ldots\ldots\ldots\ldots [28]$$

with proper interpretation being placed on the definition for α in the two cases.

Equation [28] indicates the complementary relationship between the saturation and dead-band gain factors mentioned previously. It is of significance to note that the "at-rest" condition for the two types of elements is likewise complementary. Whereas at standstill a system with a saturable element has $I_L/I_M \geq 1$, and $A_s = 1.0$ for an element with dead band $I_B/2I_M \geq 1$, and $A_D = 0$. For large values of the maximum input $I_L/I_M << 1$, A_s approaches zero as a limit and the performance of the system may depart radically from its linear operation; for $I_B/2I_M << 1$, A_D approaches unity, and the performance of the system approaches that for linear operation.

Expressed in more physical terms, for small inputs to a saturable element the system performance appears to be that for a linear system; for large inputs, the system appears to operate at greatly reduced gains. For small inputs to a system with dead band, the system is appreciably affected by the dead band and operates as if the element with dead band had reduced gain; for large inputs, the effect of the dead band appears to be small and the system tends to operate as if the element were linear.

The two sections that follow serve as illustrations of the use of these gain-change factors to obtain equivalent linear coefficients for specific systems.

SATURATION DURING SYNCHRONIZING OF POSITION CONTROL

Using linear control-system analysis, it may be shown that series or feedback stabilization methods may be used with approximately comparable effectiveness to produce a desired steady-state or transient performance. Practically, saturation effects in various portions of a control system may cause the system response to differ for the two methods of stabilization when the system is subjected to such large inputs as occur during an initial synchronizing condition of operation. Using an equivalent linear coefficient as derived in the previous section to represent the saturated element, frequency-response analyses may be made that indicate the nature of the transient response to be expected. In the following, a series-stabilized and a feedback-stabilized system, having similar saturable elements, each conditionally stable and each having approximately the same open-loop characteristic for linear operation, are compared on a frequency-response and a transient-response basis during synchronizing conditions.

For each system, a comparison is drawn between the open-loop attenuation and phase-margin versus frequency characteristics, and the closed-loop transient response for various values of the initial amplitude-ratio parameter, a_0, where

$$a_0 = \frac{\text{Limiting input to saturated element}}{\text{Maximum input to element with linear operation}} \ldots [29]$$

The "limiting input to saturated element" is the value I_L shown in Fig. 1. The "maximum input to element with linear operation" is the maximum value of the input to that element that is obtained from a transient analysis of the system using the linear values for the coefficients throughout. This linear analysis is one that is performed assuming no nonlinearities present and may be obtained using Laplace transforms or other linear-differential-equation solution methods. For a stable system the maximum value of the input to an element is its initial maximum value; hence the quantity a_0 is termed the "initial amplitude ratio" parameter. Thus, from a linear analysis of the response of a system assuming no saturation for the nonlinear element, one can determine the input to the saturating element and establish the value of the denominator for Equation [29]. From a knowledge of the saturating element's limiting value I_L, one can obtain the numerator of Equation [29].

For purposes of establishing an approximate value of the equivalent linear gain caused by an initial amplitude ratio, a_0, the gain factor versus I_L/I_M (a_0) characteristic of Fig. 15, is used with a_0 as the independent variable. Although the input to the saturated element in general will not be sinusoidal for a step input, for conditions where sustained oscillations are maintained,

the input amplitude to the saturating element tends to be approximately sinusoidal and the use of the assumed sinusoidal input is quite valid. Where sustained oscillations are not maintained and the system tends to become stable, the use of the assumed sinusoidal input indicates a value of gain for the saturated element that predicts an initial performance which compares favorably with the actual system performance. On succeeding overshoots, as they occur, the value of I_L/I_M (and gain) tends to increase and the system performance as predicted by its modified attenuation-frequency characteristic corresponds generally to the actual system response.

The phenomenon of sustained oscillations in nonlinear systems, explained on the basis of convergent and divergent equilibrium points, has been developed by Kochenburger (1) and will not be elaborated upon at this time.

Series-Stabilized System. In Fig. 16 are shown the open-loop approximate attenuation and phase-margin diagrams for a conditionally stable servo system using series-stabilizing means. The performance of this system will be considered under the conditions of saturation of the preamplifier or of the power amplifier. Shown by horizontal dotted lines is the value of the reciprocal of the gain function of the amplifier being saturated with initial amplitude ratios of $a_0 = 0.5$, 0.2, 0.05, and 0.005. These lines in effect indicate the amount by which the loop gain has been decreased by the amount of saturation noted. From Fig. 16, one notes that for $a_0 = 0.005$ a negative phase margin definitely exists at 0 decibel (db) and unstable operation in the sense of large sustained oscillations is to be expected. For $a_0 = 0.05$ the system has a small negative phase margin so that unstable operation reasonably might be expected for this condition. For values of $a_0 = 0.5$ and 0.2, the gain reduction is not sufficient to produce an unstable system. In each case, however, the speed of response of the system is decreased as the value of the frequency at which the dotted lines intersect the open-loop gain for the linear elements in the system is decreased.

Fig. 17 shows the transient response of the system of Fig. 16, synchronizing from the same position with various amounts of preamplifier saturation indicated by the values of a_0. Stable operation is indicated for all values of the initial amplitude ratio a_0, except $a_0 = 0.005$. Although the responses for the higher values of a_0 compare favorably with the linear case shown by the dotted line ($a_0 = 1.0$), for $a_0 = 0.05$ the response is quite oscillatory and for $a_0 = 0.005$ the system is definitely unstable with sustained oscillations being established.

Fig. 18 shows the synchronizing characteristics of the system in Fig. 16, with the power amplifier saturated at the values of a_0 indicated. It will be noted that with the initial amplitude ratios as high as 0.05 and 0.10, the system is unstable. For the corresponding initial amplitude ratios the same system with preamplifier saturation was stable as shown in Fig. 17. The response curves with power-amplifier saturation tend to degenerate even with a small amount of saturation, and slower responses are obtained throughout. For large amounts of saturation ($a_0 =$ 0.10 or 0.05) the characteristics indicate long periods of essentially constant-velocity operation corresponding to the saturated value of the power element. This is in contrast with the preamplifier case where the presence of stabilizing networks between the saturated element and the controlled variable prevent the latter from being so markedly influenced by the saturation of the former.

The synchronizing characteristics shown in Figs. 17 and 18 tend to bear out in a qualitative way the performance that can be predicted from the linearized approach using the equivalent linear coefficient for the saturated element. Preamplifier and power-amplifier saturation actually demonstrate a different form for their response, although this is not indicated by the linearized approach.

Feedback-Stabilized System. Fig. 19 shows the actual and the approximate open-loop attenuation and phase-margin characteristics of a conditionally stable position servo employing feedback stabilization and having similar open-loop attenuation characteristics to the series-stabilized system in Fig. 16. Saturation of the preamplifier power amplifier, and feedback amplifier will be considered separately. As indicated by the horizontal dotted lines for the values of initial amplitude ratio with preamplifier saturation, $a_0 = 0.05$ should produce unstable operation while with $a_0 = 0.5$ and 0.2 the system should be stable. Fig. 20, showing the transient synchronizing results with preamplifier saturation for these values of a_0, verifies these results.

Referring again to Fig. 19, and considering the effect of lowering the gain of the power element, it will be noted that the primary effects are to reduce the lower frequency gains, $\omega < 1.0$ radian per sec and to reduce the higher frequency gains $\omega > 20$ radians per sec. Although these effects tend to decrease the low-frequency response, they do not appreciably reduce the phase margin at 0 db gain or alter the frequency at which it occurs. On this basis it would appear that the system-synchronizing performance would not be reduced markedly by fairly large amounts of power-amplifier saturation.

Fig. 21, showing the transient synchronizing characteristics of the system in Fig. 19, with power-amplifier saturation, indicates that, for the initial amplitude ratio $a_0 = 0.05$ and greater, adequately stable operation is produced. As pointed out by consideration of the attenuation-frequency characteristics, the loss in power-amplifier gain does not affect the synchronizing characteristics of the system markedly.

Referring again to Fig. 19, consider now the effect of decreasing the feedback gain as by saturating the feedback amplifier. With decreasing amounts of feedback, the open-loop attenuation characteristic approaches that of the open loop for the system without any feedback stabilization. Since there are only two energy storages in the system without feedback, the phase margin cannot go negative, and unstable operation could not result. However, the frequency at 0 db gain is 20 radians per sec when no feedback is present and is considerably higher than this value with feedback present. Thus, for a value of initial amplitude ratio of $a_0 = 0.05$ (−24 db) for the feedback amplifier, it would appear that a lightly damped response of a frequency of about 20 radians per sec should be obtained.

Fig. 22, showing the transient response of the system with feedback saturation ($a_0 = 0.05$), while synchronizing on a fixed signal, verifies this conclusion. Although definitely oscillatory, the response is damping out and stable operation is assured. The approximate frequency of the oscillation as obtained over the first complete cycle of the transient response is slightly over 19 radians per sec and checks the linearized approach to the solution.

Conclusions Regarding Saturation During Synchronizing. From the data presented, as well as from study of other saturated systems, the following conclusions regarding system performance under conditions of saturated operation appear valid:

Saturation can be regarded as a reduction of gain of the saturated element. The saturation gain factor developed in the foregoing section provides a fairly satisfactory measure of the magnitude of the gain change produced by saturation.

For saturation effects to produce sustained oscillations, it is necessary for the system to be one in which on a linear basis a gain reduction of the saturated element can produce an unstable system.

Since for some servo systems the result of reducing the gain of one or more elements may have little detrimental effect or actually

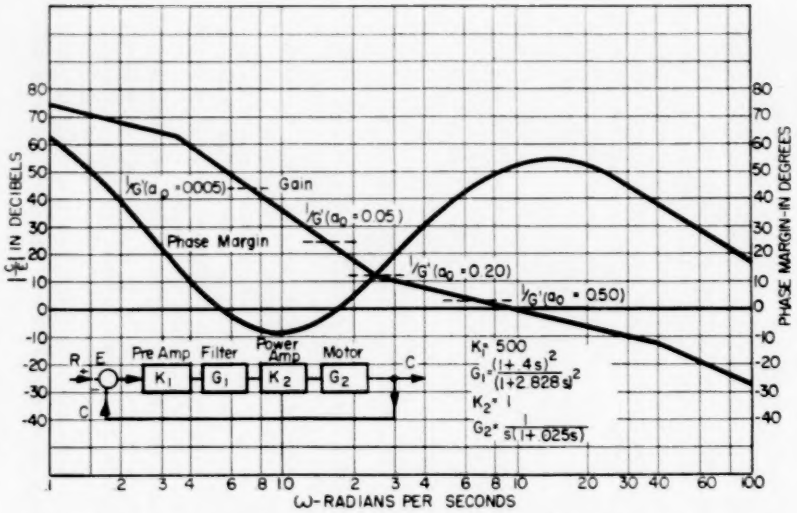


FIG. 16 OPEN-LOOP ATTENUATION (APPROXIMATE) AND PHASE MARGIN VERSUS FREQUENCY FOR SERIES-STABILIZED POSITION-CONTROL SYSTEM
(Gain reduction caused by initial amplitude ratio as noted.)

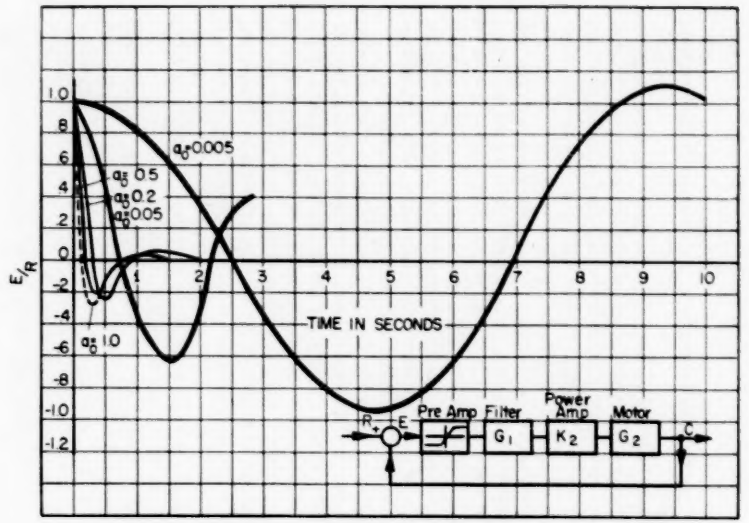


FIG. 17 SYNCHRONIZING ON A FIXED SIGNAL WITH PREAMPLIFIER SATURATION
(Series stabilization as in Fig. 16, with initial amplitude ratios as noted.)

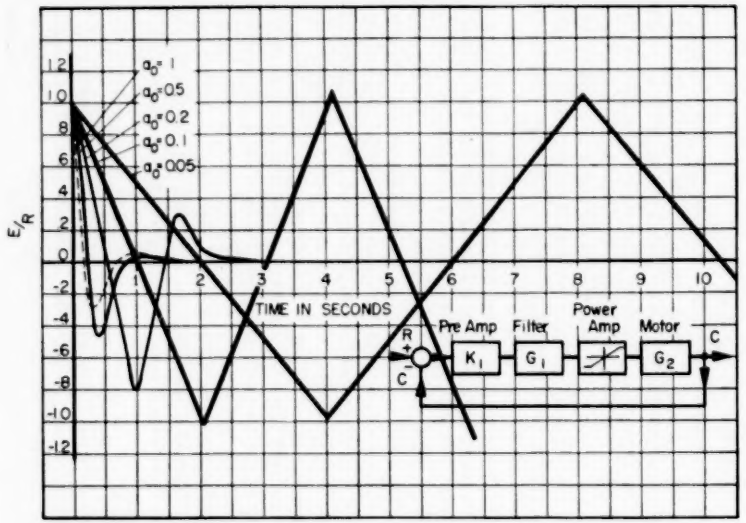


FIG. 18 SYNCHRONIZING ON A FIXED SIGNAL WITH POWER-AMPLIFIER SATURATION
(Series stabilization as in Fig. 16, with initial amplitude ratios as noted.)

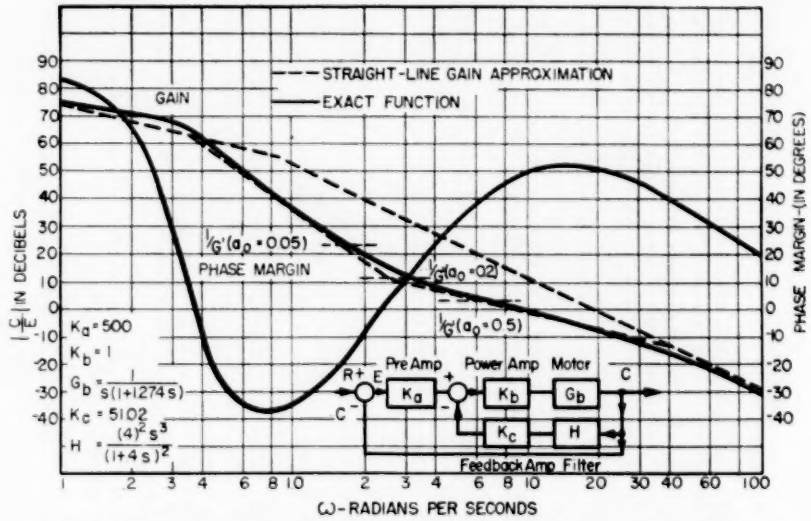


FIG. 19 OPEN-LOOP ATTENUATION AND PHASE MARGIN VERSUS FREQUENCY FOR FEEDBACK-STABILIZED POSITION-CONTROL SYSTEM
(Preamplifier gain reduction caused by initial amplitude ratio as noted.)

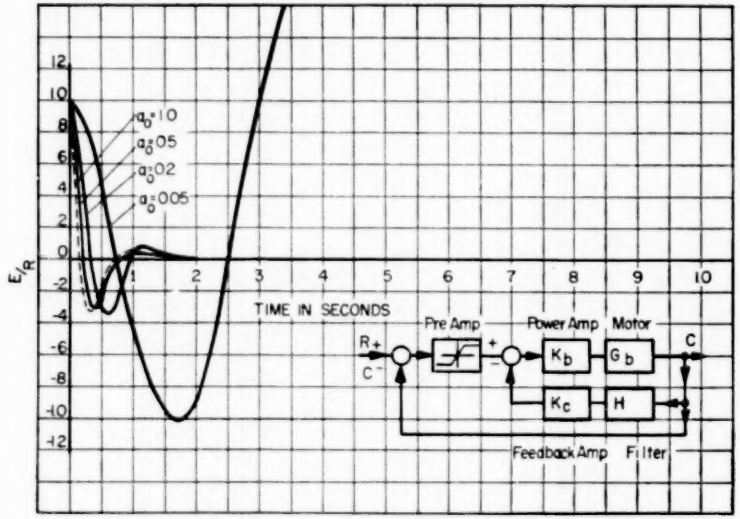


FIG. 20 SYNCHRONIZING ON A FIXED SIGNAL WITH PREAMPLIFIER SATURATION
(Feedback stabilization as in Fig. 19, with initial amplitude ratios as noted.)

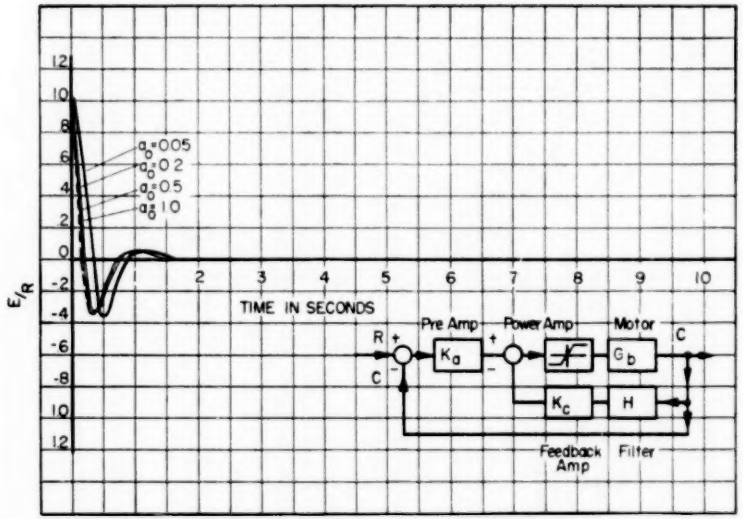


FIG. 21 SYNCHRONIZING ON A FIXED SIGNAL WITH POWER-AMPLIFIER SATURATION
(Feedback stabilization as in Fig. 19, with initial amplitude ratios as noted.)

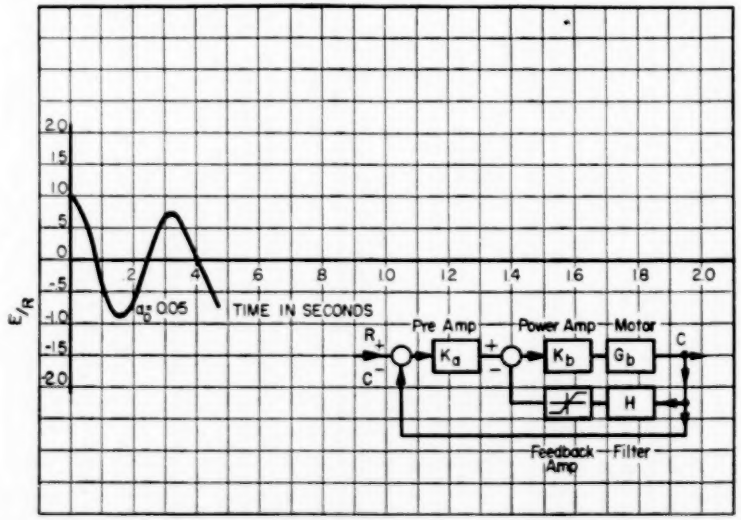


FIG. 22 SYNCHRONIZING ON A FIXED SIGNAL WITH FEEDBACK-AMPLIFIER SATURATION
(Feedback stabilization as in Fig. 19, with a_0 = 0.05.)

may have even a somewhat salutary effect, it appears that saturation may not always produce unsatisfactory performance.

These conclusions are in good agreement with the findings of Kochenburger (3) and others (4).

Effect of Backlash on Control-System Mechanical Design

In the manufacture of position-control systems there is a limit to the degree of perfection by which the motor and load may be connected. Generally the motor is geared to the load and there exists some backlash in the gear mesh. A number of detailed analyses (5, 6) have been presented which describe the performance of the motor and load with backlash present and with various types of load characteristics assumed. In a companion paper by C. H. Thomas (7) there is a comprehensive treatment of the subject of dead band which deals with a number of different backlash conditions.

It is also of interest to approach this problem as an extension of present-day servo-analysis methods. In this material, emphasis will be placed on using the describing function technique after obtaining an equivalent linear coefficient for the spring-backlash combination. With this as a tool, the effect of the ratio of the inertia in the motor to the inertia in the load, J_M/J_L, the effects of mechanical resonance c, and damping ratio ζ, of the load itself, and the effect of load versus motor-position feedback will be investigated by frequency-response methods. The qualitative results thus obtained will be compared with quantitative results of analog-computer studies of these same cases.

To provide useful data for mechanical designers of servo-equipment, the effect of backlash in a common form of position control, as shown in Fig. 23, will be considered when subjected to small step-input disturbances. The linear elements for the system, preamplifier, stabilizing networks, and power amplifier, are shown by the blocks. The motor and load are shown schematically and correspond to these same elements considered previously in Fig. 8. A block diagram for these elements was developed from Equations [13] and [16] and is shown in Fig. 11. Using an equivalent linear spring constant K_s' to represent the spring-backlash combination and the block diagram in Fig. 11, one can express the relationship between the motor position θ_M and the input voltage to the motor I_N as

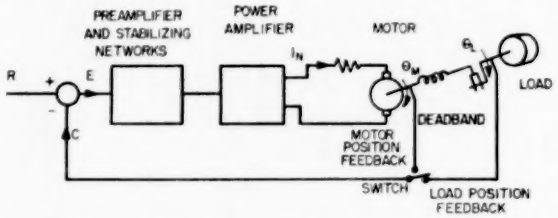


Fig. 23 Position Control Having Mechanical Dead Band and Spring-Mass, Viscous-Damping Load

$$\frac{s\theta_M}{I_N} = \frac{K_T/R\,(J_L s^2 + D_L s + K_s')}{J_M J_L s^3 + \left(J_M D_L + J_L \frac{K_s K_T}{R}\right) s^2 + \left[(J_M + J_L)K_s' + D_L \frac{K_s K_T}{R}\right] s + K_s'\left(D_L + \frac{K_s K_T}{R}\right)} \quad \ldots\ldots [30]$$

In similar fashion the ratio of load position θ_L, to the input to the motor I_N, is

$$\frac{s\theta_L}{I_N} = \frac{K_s'\,K_T/R}{J_M J_L s^3 + \left(J_M D_L + J_L \frac{K_s K_T}{R}\right) s^2 + \left[(J_M + J_L)K_s' + D_L \frac{K_s K_T}{R}\right] s + K_s'\left(D_L + \frac{K_s K_T}{R}\right)} \quad \ldots\ldots [31]$$

These equations are identical with the corresponding ones for these elements without backlash with the exception that the spring constant K_s' is a function of the amplitude of $\theta_B/2$ and the amplitude difference maximum $(\theta_M - \theta_L)_{max}$; K_s' may vary from 0 for $-\theta_B/2 < \theta_M - \theta_L < +\theta_B/2$ to K_s the actual shaft stiffness gradient for $|\theta_M - \theta_L| >> |\theta_B/2|$. Equations for motor and load positions have been obtained separately so that a comparison can be made of the relative merits of motor position versus load position as the controlled variable for feedback purposes, see switch in Fig. 23.

System-Design Parameters. To provide a starting point for this analysis on a nonlinear basis, the system will be considered as a simple linear system and the mechanical design parameters will then be expressed in terms of the characteristics of this linear system. It will be assumed that the motor and load inertia, J_M and J_L, respectively, can be lumped into one equivalent inertia referred to the motor shaft. The stiffness coefficient of the load is assumed infinite and its viscous and coulomb damping are considered negligible. For this case, Equations [30] and [31] become equal and contain only the single "motor" time-constant term

$$\frac{s\theta_M}{I_N} = \frac{s\theta_L}{I_N} = \frac{1}{K_s\left[\frac{(J_M + J_L)R}{K_s K_T} s + 1\right]} \quad \ldots\ldots [32]$$

where

$K_s = 2.00$ volts/rad/sec

$\frac{K_T}{R} = 1.25$ lb-ft/volt

$J_M + J_L = 0.25$ slug ft^2

$T_T = \frac{(J_M + J_L)R}{K_s K_T} = 0.1$ sec, total equivalent motor time constant

The transfer function of the power amplifier is

$$G_2 = \frac{20}{\left(\frac{s}{50} + 1\right)} \quad \ldots\ldots [33]$$

The preamplifier and stabilizing networks have been so chosen that, with the assumptions noted, the system is conditionally stable. This transfer function is

$$G_1 = \frac{40\left(\frac{s}{0.64} + 1\right)^2}{\left(\frac{s}{0.064} + 1\right)^2} \quad \ldots\ldots [34]$$

Fig. 24 shows the open-loop approximate attenuation and phase characteristics as a function of frequency in the vicinity of 0 db for this linear system.

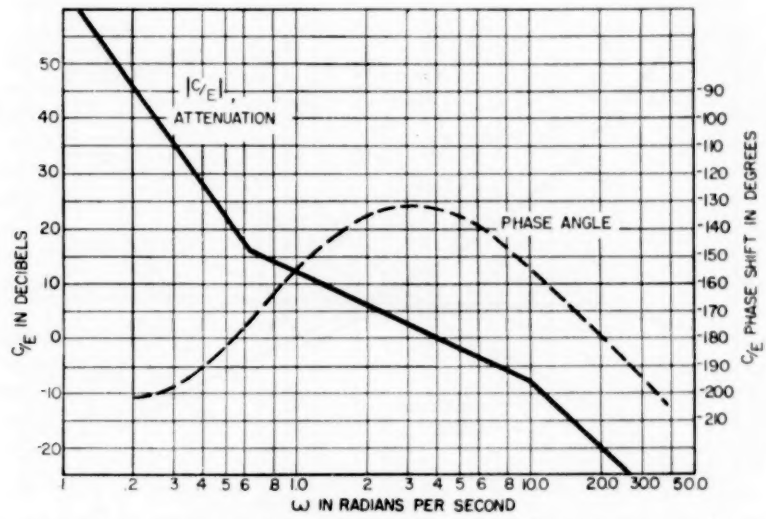


FIG. 24 OPEN-LOOP APPROXIMATE ATTENUATION AND PHASE-SHIFT CHARACTERISTICS FOR LINEARIZED SYSTEM CORRESPONDING TO FIG. 23

To permit the mechanical parameters of the load and motor to be entered into the problem in a systematic fashion, the following definitions are employed:

$$\left.\begin{aligned}
\omega_0 &= \sqrt{\frac{K_s}{J_L}} && \text{undamped natural frequency of load alone with motor inertia fixed (mechanical resonance frequency)}\\
\omega_T &= \frac{1}{T_T} && \text{frequency corresponding to reciprocal of total equivalent motor time constant}\\
c &= \frac{\omega_0}{\omega_T} && \text{ratio of load mechanical resonance to equivalent motor time-constant frequency}\\
\zeta &= \frac{D_L}{2J_L\omega_0} && \text{damping ratio for mechanical load alone with motor fixed and no backlash}\\
A_D &= \frac{K_s'}{K_s} && \text{spring-constant reduction ratio, and expressible in terms of } \frac{\theta_B}{2(\theta_M-\theta_L)_{max}}\\
& && \text{by means of Fig. 15}
\end{aligned}\right\} \quad \ldots [35]$$

Further it is assumed that the combined motor and load inertia $(J_M + J_L)$ is a constant equal to 0.25 slug-ft², although the ratio of J_M/J_L is a parameter that the servo designer can control.

With the aid of these definitions and assumptions, the mechanical design parameters can be expressed in a somewhat nondimensional form and the effects of changes in these ratios may be investigated analytically in a systematic fashion.

Basis for Stability Analysis. Kochenburger has outlined a method for separating a nonlinear system into a linear portion having a transfer function and a nonlinear portion having a describing function (1). The system under consideration has been so represented in block diagram form in Fig. 25. For convenience, the motor integration is included with the linear transfer function. From the closed-loop response

$$\frac{C}{R} = \frac{G(\omega)\,G'(A_D, \omega)}{1 + G(\omega)\,G'(A_D, \omega)} \quad \ldots [36]$$

The conditions for system instability can be expressed as

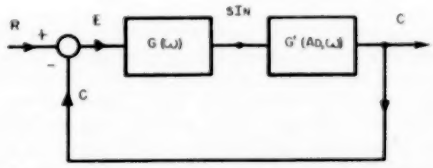


FIG. 25 SYSTEM OF FIG. 23, REPRESENTED AS A SERIES COMBINATION OF TRANSFER AND DESCRIBING FUNCTIONS

$$G(\omega) = -\frac{1}{G'(A_D, \omega)} \quad \ldots [37]$$

where $$G(\omega) = \frac{800\,(s/0.64 + 1)^2}{s(s/0.064 + 1)^2\,(s/50 + 1)}$$

and $G'(A_D, \omega)$ is the describing function given by Equation [30] or [31], depending on whether load or motor-position feedback is employed, and is a function of the amplitude ratio

$$\alpha = \frac{\theta_B}{2(\theta_M - \theta_L)_{max}}$$

as well as frequency.

To facilitate the use of the attenuation-frequency techniques, the right and left sides of Equations [37] are plotted for different assumed values of A_D, for each of the conditions of system parameters that are of interest. At those frequencies where the magnitude equality of Equation [37] is realized, the phase-angle relationship which exists is noted. If this phase relationship is such that the phase margin is positive for the closed loop, the system is stable for this condition of A_D and the magnitude of the oscillation initially present tends to decrease such that a smaller value of A_D is realized. If the phase margin is negative at the frequency where the magnitude equality of Equation [37] exists, the system is unstable for this condition of A_D and the magnitude of the oscillation initially present tends to increase such that a larger value of A_D is realized.

When a condition exists where both the magnitude and angle relationships of Equation [37] occur simultaneously, the oscillation is sustained at this gain-factor change (A_D) and sustained oscillation frequency ω_s. From the gain-factor change and the

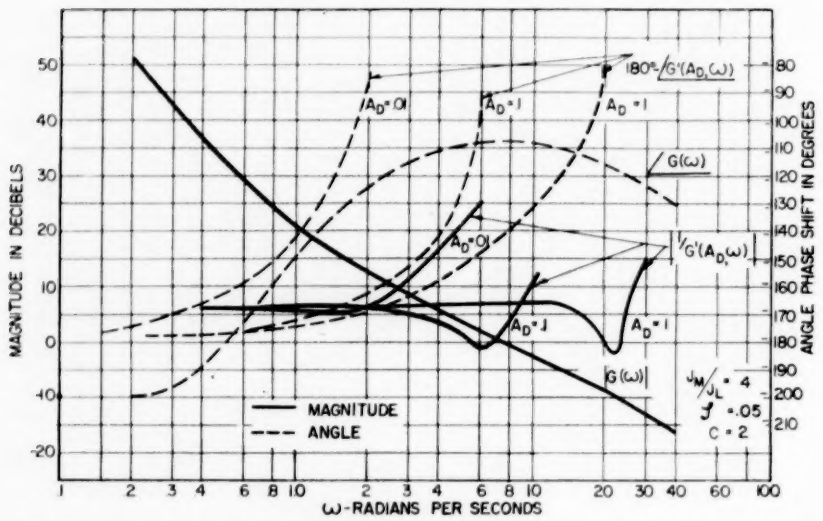


FIG. 26 ATTENUATION AND PHASE CHARACTERISTICS OF TRANSFER FUNCTION AND NEGATIVE RECIPROCAL OF DESCRIBING FUNCTION OF FIG. 25, SHOWING EFFECT OF DEAD-BAND GAIN FACTOR A_D WITH LOAD FEEDBACK

curve of A_D versus $I_B/2I_M$ given in Fig. 15, one can determine the value of the backlash-input maximum ratio. Knowing the magnitude of the backlash angle for the given control, and the $\theta_B/2(\theta_M - \theta_L)_{max}$ ratio from the curve, the value for $(\theta_M - \theta_L)_{max}$ can be calculated. With this the magnitudes of other significant quantities such as error, or load position may be determined at the frequency of oscillation ω_s using regular linear-analysis methods.

To illustrate the stability aspect of this process, refer to Fig. 26, in which the magnitude and phase angle of $G(\omega)$ of Equation [37] are shown by solid and dotted lines, respectively. The magnitude and phase angle of $-1/G'(A_D, \omega)$ also are shown by solid and dotted lines, respectively, for values of $A_D = 1.0$, 0.1, and 0.01. Load-position feedback is employed, since a greater tendency for sustained oscillations is noted when load rather than motor-position feedback is used. For $A_D = 1.0$, corresponding to the linear case in which no backlash is present, the equality of magnitudes for Equation [37] occurs at $\omega = 3.7$ radians per sec although a positive phase margin exists until $\omega = 15$ radians per sec. Hence stable operation can exist for the values of system and load parameters indicated.

With $A_D = 0.1$, the equality of magnitudes for Equation [37] occurs at $\omega = 7.0$ radians per sec, although a negative phase margin exists for frequencies greater than $\omega = 5.5$ radians per sec. This means that the system is unstable for this value of A_D, corresponding to $\theta_B/2(\theta_M - \theta_L)_{max} = 1.24$ and the oscillations will increase to a larger amplitude before stable sustained oscillations are reached. The magnitude and phase curves for $A_D = 0.01$ also indicate an unstable condition so that the possibility definitely exists for starting up the oscillation process that will grow to produce larger values of A_D than 0.1.

This method for determining whether sustained oscillations can exist in the region from $0.1 < A_D < 1.0$ is used in the material that follows. This range of A_D has been chosen since it is felt that it may be possible to tolerate values of sustained oscillations of less than 1.25 times the value of backlash. Further, the use of this method will serve to produce a qualitative indication of the relative effects of the various mechanical design parameters that are of principal interest.

Effect of Load Mechanical-Resonant Frequency. In Fig. 27 is shown the effect of increasing the load mechanical resonant-frequency ratio from $c = 1$ to 4 for the mechanical design parameters $J_M/J_L = 1.0$ and $\zeta = 0.1$ with load-position feedback. Study of these curves indicates that for $A_D = 0.1$ and $c = 1.0$, the system is unstable while for the higher values of c and $A_D = 0.1$ and for all values of c with $A_D = 1.0$ the system is stable. Further, the relative degree of stability improves with increasing values of c, as indicated by a greater amount of phase margin at magnitude equality and by the higher ratio of frequencies between magnitude equality and phase-margin zero. These curves serve to indicate then that increasing the load mechanical resonance-frequency ratio reduces the magnitude of the sustained oscillation caused by backlash. Values of $c = 2$ or greater appear to be satisfactory for the system here considered.

Effect of Motor-to-Load Inertia Ratio, J_M/J_L. Fig. 28 shows the effect of varying the motor-to-load inertia ratio (J_M/J_L) over the range $1/4$ to 4 on the stability of the system using load-position feedback. At $A_D = 0.1$ and $J_M/J_L = 1/4$ the system is shown to be stable. However, for values of $J_M/J_L = 1$, and $J_M/J_L = 4$, and $A_D = 0.1$ the system is shown to be unstable with increasing instability for the higher value of J_M/J_L. At $A_D = 1.0$ all three values of J_M/J_L indicate stable systems with the lower the value of J_M/J_L, the more desirable the performance. Therefore, for $J_M/J_L = 1$ and 4, sustained oscillations of greater than 1.25 times the backlash magnitude will be established. In this case where load-position feedback is being employed and the total inertia is constant, it is desirable to have as much as possible of the total load inertia to be located in the load for minimizing the oscillations caused by backlash.

Effect on Load-Damping Ratio, ζ. Fig. 29 shows the effect of varying the damping ratio of the load ζ, over the range from 0.025 to 0.10 for values of $A_D = 0.10$ and 1.0 with load-position feedback. With $\zeta = 0.10$ and $A_D = 0.1$, the system appears to be nearly on the limit of stability, although for values of $\zeta = 0.05$ and 0.025 the system can be seen to be definitely unstable. For $A_D = 1.0$, the system is shown to be stable for all three values of damping ratio, although the lower the damping ratio, the less stable the system in the range of frequencies near the mechanical resonance frequency of the load itself. Thus considered on the basis of minimizing the magnitude of the sustained oscillation

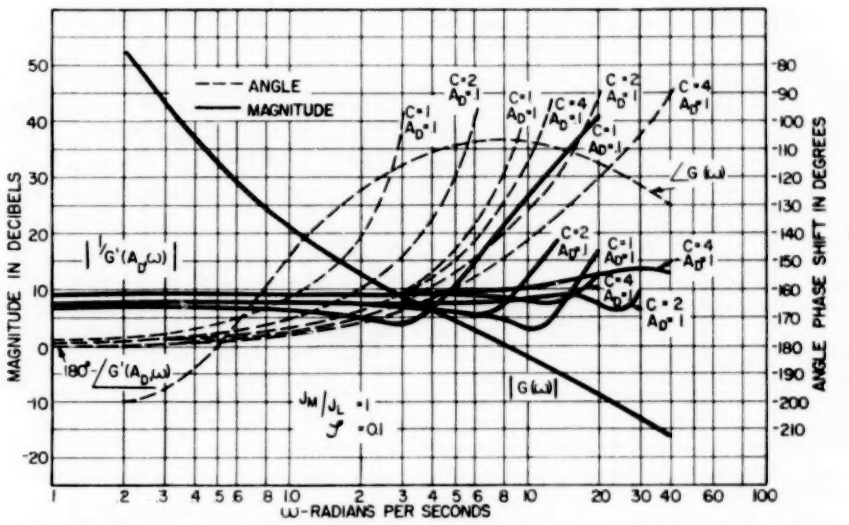


FIG. 27 ATTENUATION AND PHASE CHARACTERISTICS OF TRANSFER FUNCTION AND NEGATIVE RECIPROCAL OF DESCRIBING FUNCTION OF FIG. 25, SHOWING EFFECT OF MECHANICAL RESONANCE-FREQUENCY RATIO c, WITH LOAD FEEDBACK FOR $A_D = 0.1$ AND 1.0

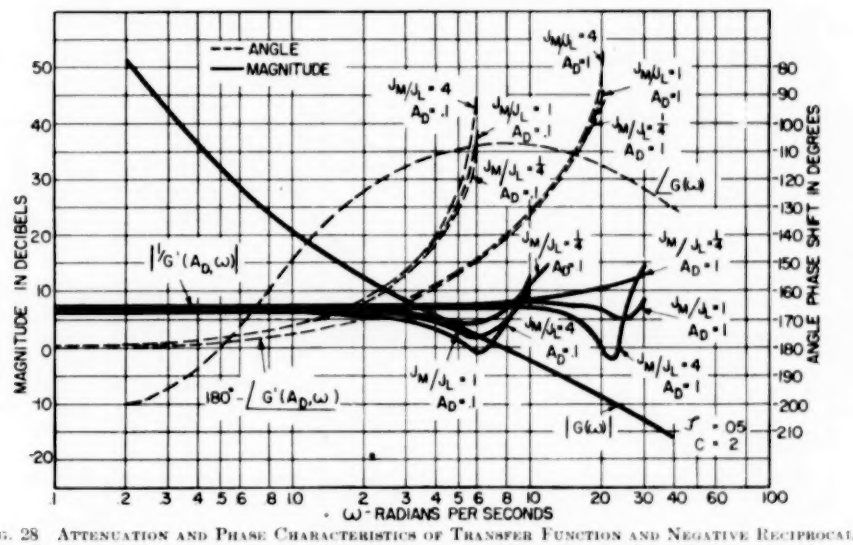


FIG. 28 ATTENUATION AND PHASE CHARACTERISTICS OF TRANSFER FUNCTION AND NEGATIVE RECIPROCAL OF DESCRIBING FUNCTION OF FIG. 25, SHOWING EFFECT OF MOTOR INERTIA/LOAD INERTIA (J_M/J_L) WITH LOAD FEEDBACK FOR $A_D = 0.1$ AND 1.0

resulting from backlash, it is desirable to have a mechanical damping ratio in the vicinity of 0.10. Since viscous damping represents a power loss for constant-velocity load motion, it is apparent that a limited amount of viscous damping should be used and other means should be sought for eliminating the effects of sustained oscillations resulting from backlash.

Comparison of Load and Motor Feedback. Fig. 30 shows a comparison of the stability of a given system using either motor feedback or load feedback as shown in Fig. 23, for $A_D = 0.10$ and $A_D = 1.0$. For the case of load feedback, definite instability is indicated at $A_D = 0.10$, while stable operation is shown at $A_D = 1.0$. Sustained oscillations of greater than 1.25 times the value of the backlash angle are thus predicted with load feedback. For the same system but with motor feedback, relatively highly stable operation is indicated for both values of A_D. This result confirms the general experimental findings of servo designers over the years. Further, it serves to indicate that the additional phase shift produced by using load-position feedback rather than motor-position feedback, Equations [30] and [31], may be objectionable over the range of frequencies of interest. Although at low frequencies, $\omega < 1.0$ radian per sec, there is not much difference between the results obtained from two sources of position-feedback signals, above this frequency range the results vary quite markedly. Viewed on this basis, the "divided-reset" methods recommended and used by the British and others appear to have particular merit. These results confirm the long-established design criterion, that "whenever possible the position response to be used for feedback pur-

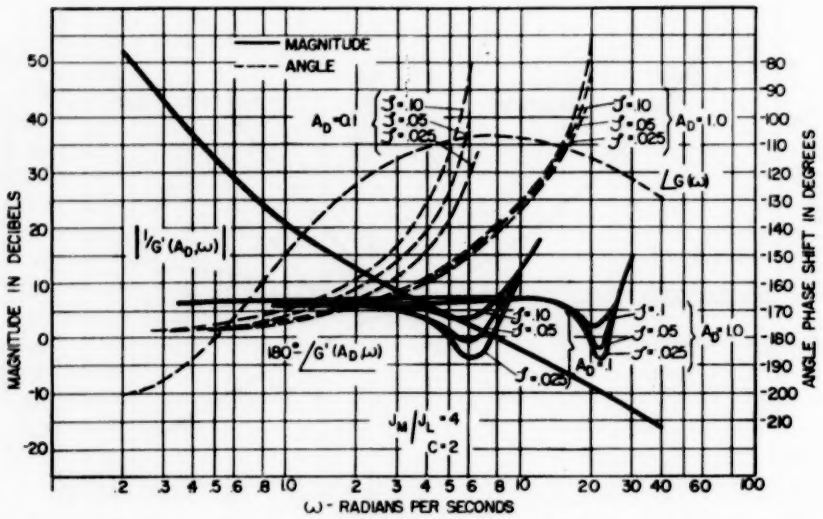


FIG. 29 ATTENUATION AND PHASE CHARACTERISTICS OF TRANSFER FUNCTION AND NEGATIVE RECIPROCAL OF DESCRIBING FUNCTION OF FIG. 25, SHOWING EFFECT OF MECHANICAL DAMPING RATIO ζ WITH LOAD FEEDBACK FOR A_D = 0.1 AND 1.0

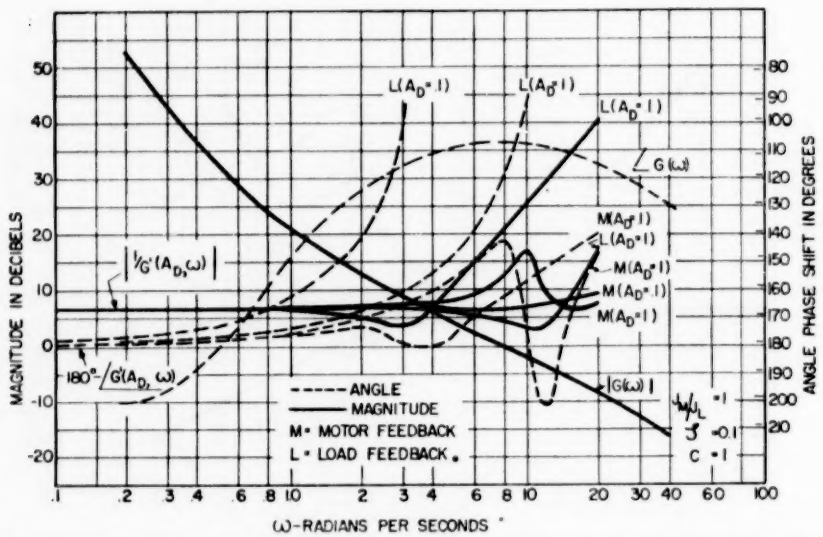


FIG. 30 ATTENUATION AND PHASE CHARACTERISTICS OF TRANSFER FUNCTION AND NEGATIVE RECIPROCAL OF DESCRIBING FUNCTION OF FIG. 25, SHOWING EFFECT OF MOTOR VERSUS LOAD FEEDBACK FOR A_D = 0.1 AND 1.0

poses should be obtained from the motor rather than the load for most stable operation."

Analog-Computer Results. The position-control system shown in Fig. 23 has been set up on an electronic analog computer by R. L. Mathews, and its performance has been determined for a wide variety of system parameters with load or motor feedback signals. Figs. 31, 32, and 33 show for load feedback a number of significant performance characteristics as a function of the mechanical damping ratio for three different values of the motor/load inertia ratio J_M/J_L. The performance characteristics that are indicated include the frequency of the sustained oscillation ω_s, the amplitude of the error to the backlash ratio $2E/\theta_B$, and the amplitude of the maximum of the twist angle to the backlash ratio $2(\theta_M - \theta_L)_{max}/\theta_B$. The latter two characteristics are merely amplitude ratios and the magnitudes of the error or twist angles that are obtained are directly proportional to the magnitude of the backlash angle θ_B. From this it is apparent that the effects of backlash can be minimized by limiting the amount of the backlash angle.

Referring to Fig. 31, for which $J_M/J_L = 1/4$, i.e., the majority of the inertia is in the load, sustained oscillations are shown to exist for load mechanical resonance-frequency ratios of $c = 1$ and 2, with no sustained oscillations existing for $c = 4$. The sustained-oscillation frequencies are in the rather limited range from 3 to 6 radians per sec, i.e., in the vicinity of the open-loop crossover in Fig. 24. The error/backlash and twist angle/

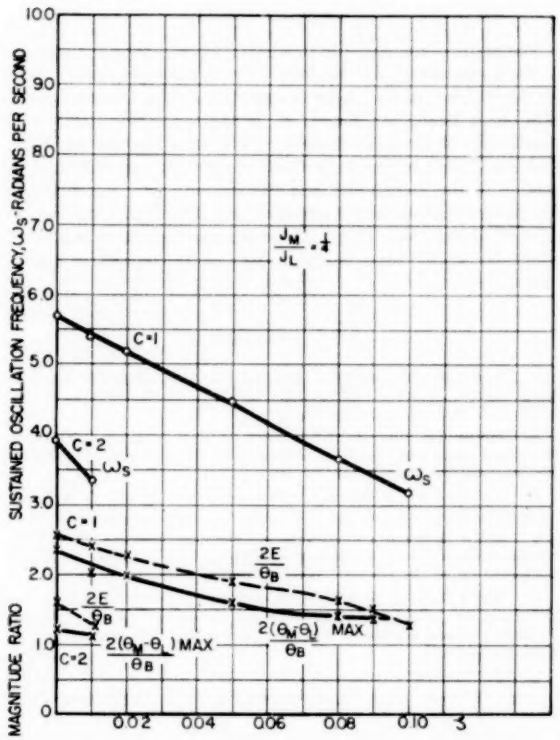


FIG. 31 FREQUENCY OF SUSTAINED OSCILLATION ω_s, (MOTOR-LOAD DIFFERENCE ANGLE/BACKLASH ANGLE) AND (ERROR ANGLE/BACKLASH ANGLE) AS A FUNCTION OF DAMPING RATIO FOR VARIOUS VALUES OF LOAD MECHANICAL RESONANCE RATIO c WITH $J_M/J_L = 1/4$

backlash ratios are approximately the same and vary from 2.5 to about 1.0 with the magnitudes decreasing gradually with increasing values of the damping ratio ζ. For increasing values of c, the amplitudes of the sustained oscillations are smaller, and, for $c = 4$ and greater, the oscillations fail to be sustained.

Fig. 32 shows similar data to those presented in Fig. 31, but for the case of $J_M/J_L = 1.0$. For this case in which the motor and load each have the same equivalent inertia, sustained oscillations exist with load mechanical resonance-frequency ratios of 1, 2, and 4. The frequency range of the sustained oscillation is greater extending from 2.0 to 7.0 radians per sec with the oscillation frequency for low damping ratios being greater than the open-loop crossover frequency. For the same damping ratio and c, the error and twist angle/backlash ratios are greater with $J_M/J_L = 1$ than with $J_M/J_L = 1/4$. However, with $c \geq 2$ these ratios are less than 2.0 and represent less than doubling the value of the backlash itself. As in the previous case, increasing the load-damping ratio causes the magnitude and frequency of the error and twist oscillations to decrease.

Fig. 33, for which $J_M/J_L = 4$, indicates greatly reduced performance for the same load mechanical resonance frequencies and the same mechanical damping ratios as for the lower values of J_M/J_L. With $c = 1.0$ it is impossible to decrease the damping ratio below 0.03 without producing oscillations of increasing amplitudes. Even for higher values of ζ (up to 0.1) prohibitively high values of error and twist angle to backlash ratios are produced. For $c \geq 2$, despite the fact that fairly high-frequency sustained oscillations are produced, the error and twist to backlash angle-amplitude ratios are again less than 2.0 with $J_M/J_L = 4.0$.

COMPARISON OF ANALOG-COMPUTER AND FREQUENCY-RESPONSE ANALYSIS RESULTS

Comparing the analog-computer results with the equivalent linearized-coefficient results, it appears that the same qualitative conclusions are indicated by each for the effects of c, J_M/J_L, and ζ on the sustained oscillations caused by backlash. However, in general the sustained-oscillation frequency ω_s is from 10 to 40 per cent higher when obtained from the linearized approach than when determined by the analog computer. There is a somewhat better correlation of the oscillation amplitudes obtained by the two methods, however.

The analog-computer traces show that for most conditions of sustained oscillation, the difference angle $\theta_M - \theta_L$ is dominantly of fundamental frequency. As such, these traces provide an independent check of the assumption that there is a sinusoidal input to the backlash function. It will be recalled that this assumption forms the basis of the equivalent linear-coefficient approach to this method of analysis of systems with backlash and dead band.

Studies, using motor-position feedback for both computer and frequency-response analyses, showed stable operation with no sustained oscillations for the cases studied.

Summary of Design Results of Backlash Study. Computer and linearized frequency-response results indicate for the conditions of study that sustained oscillations caused by backlash can be reduced by the following:

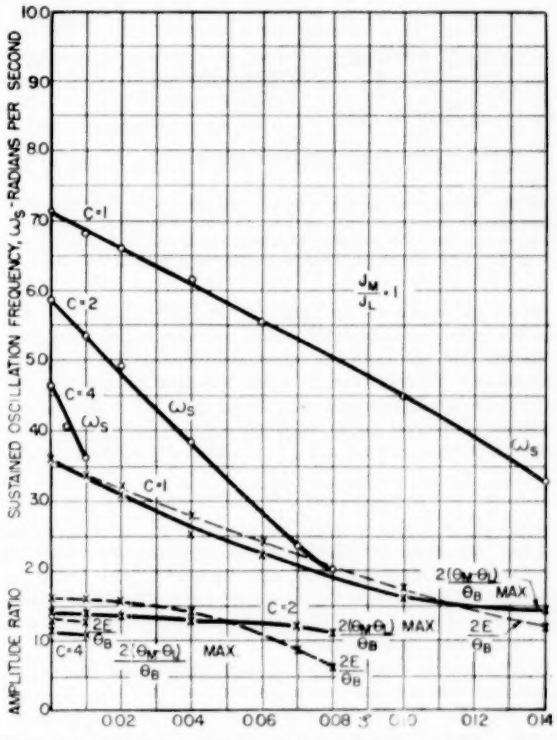


FIG. 32 FREQUENCY OF SUSTAINED OSCILLATION ω_s, (MOTOR-LOAD DIFFERENCE ANGLE/BACKLASH ANGLE) AND (ERROR ANGLE/BACKLASH ANGLE) AS A FUNCTION OF DAMPING RATIO FOR VARIOUS VALUES OF LOAD MECHANICAL RESONANCE RATIO c WITH $J_M/J_L = 1$

(*a*) Increasing the load mechanical resonance frequency c.
(*b*) Decreasing the ratio of motor inertia to load inertia J_M/J_L.
(*c*) Increasing the load-viscous damping ratio ζ.

Use of motor rather than load-position feedback also is beneficial in reducing oscillations caused by backlash.

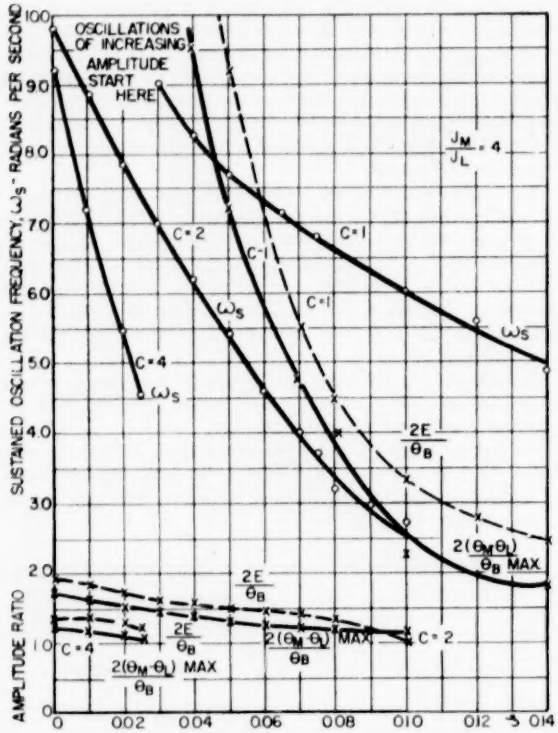


FIG. 33 FREQUENCY OF SUSTAINED OSCILLATION ω_S, (MOTOR-LOAD DIFFERENCE ANGLE/BACKLASH ANGLE) AND (ERROR ANGLE/BACKLASH ANGLE) AS A FUNCTION OF DAMPING RATIO FOR VARIOUS VALUES OF LOAD MECHANICAL RESONANCE RATIO c WITH $J_M/J_L = 4$

Although the results reported here are limited to the case of a conditionally stable position control, similar results also were obtained and are valid for an unconditionally stable system. Reference to Figs. 26 through 29 indicates that the phenomenon of sustained oscillation takes place in the range of frequencies near open-loop crossover and is not affected appreciably at the much lower frequencies corresponding to conditionally stable operation.

CONCLUSIONS

This paper indicates there are numerous nonlinear elements having saturation or dead-band characteristics for which an "equivalent linear coefficient" may be used as the describing function. The equivalent gain factors for saturation and dead band are shown to have a somewhat complementary nature and to be of relatively simple form.

Although the results of using this approximate method of analyzing backlash and saturation do not provide an exact solution, they do permit the designer to obtain qualitative results. As such, this method is particularly helpful for use in the more complex systems such as generally are encountered in practice, but that do not lend themselves to more exact analytical treatment.

ACKNOWLEDGMENTS

The material presented in this paper represents only a portion of more complete data obtained by a number of people. The author wishes to acknowledge the assistance of Mrs. Cecile Lester, Mrs. Jean Bean, Mr. Ernest H. Bennett, and Mr. Richard L. Mathews in obtaining these data. Miss Anna Kosinski and Mrs. Dolores Riccio are also to be thanked for their part in the preparation of the manuscript and the figures. The comments of Mr. Warren Gaines on the material contained in this manuscript were of considerable help. The encouragement and assistance which Dr. C. F. Green has given this effort are greatly appreciated.

BIBLIOGRAPHY

1 "A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, 1950, pp. 270–284.

2 "Sinusoidal Analysis of Feedback Control Systems Containing Nonlinear Elements," by E. C. Johnson, AIEE Technical Paper No. 52–154, 1952 (not published).

3 "Limiting in Feedback Control Systems," by R. J. Kochenburger, AIEE Technical Paper No. 53–109, 1953 (not published).

4 "Some Saturation Phenomena in Servomechanisms With Emphasis on the Tachometer Stabilized System," by E. Levinson, AIEE Technical Paper No. 53–110, 1953 (not published).

5 "The Effects of Backlash and of Speed Dependent Friction on the Stability of Closed-Cycle Control Systems," by A. Tustin, *Journal of the Institute of Electrical Engineers*, vol. 94, part II-A, 1947, pp. 143–151.

6 "Backlash and Resilience Within the Closed-Loop of Automatic Control Systems," by J. H. Liversidge, *Automatic Control*, Academic Press, New York, N. Y., edited by A. Tustin, 1952, pp. 343–375.

7 "Stability Characteristics of Closed-Loop Systems With Dead Band," by C. H. Thomas, published in this issue, pp. 1365–1382.

Discussion

W. M. GAINES.[3] This paper shows the value of the "describing-function" method of analysis when dealing with a number of practical, nonlinear, engineering problems. The availability of analog computers in many cases has reduced the need for hand calculations from one of obtaining exact quantitative results to one of getting a qualitative understanding of the effects. It is under these conditions that the describing-function method of analysis plays a particularly important role. Although the method is not accurate, it permits rapid evaluation of a system's performance and it allows extrapolation and interpolation of computer results.

There are two comments the writer would like to make which are in the way of additions to the paper and not criticism. The point of accuracy should be stressed. Those who are familiar with the frequency-response method in design of servomechanisms and regulators have come to rely heavily upon the accuracy of the results obtained from this approach. The same accuracy cannot be expected from the describing-function analysis. This is illustrated graphically by the following simple example:

Consider the second-order system with backlash shown in Fig. 34 of this discussion. This system can be analyzed exactly.[4,5] The characteristic of the backlash is shown in Fig. 35 and is of the

[3] Development Engineer, Aero. and Ordnance Systems Division, General Electric Company, Schenectady, N. Y.

[4] "Instrument Inaccuracies in Feedback Control Systems With Particular Reference to Backlash," by H. T. Marcy, M. Yachter, and J. Zauderer, AIEE Technical Paper 49-197.

[5] "Dynamics of Automatic Control," by R. C. Oldenbourg and H. Sartorius, Trans. ASME, vol. 70, 1948, pp. 168–173.

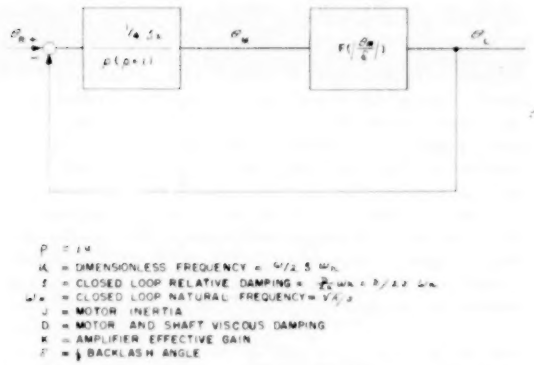

P = 1/K
K = DIMENSIONLESS FREQUENCY = ω/2 δ ω_n
ζ = CLOSED LOOP RELATIVE DAMPING = P/2 ω_n = D/2 J ω_n
ω_n = CLOSED LOOP NATURAL FREQUENCY = √K/J
J = MOTOR INERTIA
D = MOTOR AND SHAFT VISCOUS DAMPING
K = AMPLIFIER EFFECTIVE GAIN
δ = ½ BACKLASH ANGLE

FIG. 34 GENERALIZED SECOND-ORDER SERVO WITH BACKLASH

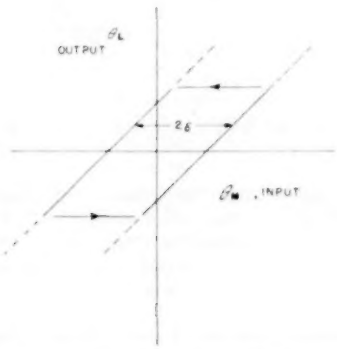


FIG. 35 HYSTERESIS-TYPE BACKLASH—FRICTION LOAD

simple hysteresis type. The describing function for such a nonlinearity is shown in Fig. 36, herewith, along with the frequency locus of the servo, which results in two intersections with the describing function. After the nomenclature of Kochenburger,[6] the first intersection would be called a divergent point and the second intersection a convergent point (equivalent to a limit cycle in the phase plane). From this plot it would appear that unless disturbed sufficiently to exceed the divergent equilibrium point, the system would operate in a stable manner.

Several authors[4,5] have shown by means of exact analysis that as long as a system of this type is underdamped there will exist oscillations of some finite amplitude. These same analyses have shown that no region exists where the system exhibits a stable limit cycle and an unstable limit cycle for the same system constants. In justification of the describing-function analysis, however, it should be pointed out that the divergent equilibrium intersection occurred at a magnitude which is of little practical importance, and that the frequency and magnitude of oscillation predicted from the convergent point check very closely with the actual values obtained.

In the foregoing example where the predicted amplitude of oscillation was 2δ and the frequency was 1.6 radians per sec, analog-computer results have shown a peak amplitude of oscillation of 2δ and a fundamental frequency of oscillation of 1.5 rps.

The author's analysis of the imperfect coupling of a motor and load is particularly timely. There is a growing interest, brought about by the need for higher performance, in this problem of representing the load resonance and backlash in power servos. In such servos the resonance and backlash effects often impose the limitation on the maximum servo performance obtainable. The author indicates the desirability of having a load relative damping of at least 0.1. However, in most practical systems, where the load inertia is the same order of magnitude as the motor, the load relative damping factor will be 5 to 20 times smaller than the figure quoted. In many cases the necessary damping is actually obtained from the static and coulomb friction present.

It may be pointed out that the oscillations indicated by the author's analysis were all of the high-frequency type occurring near the crossover of the system. In actual practice, a low-frequency oscillation (commonly referred to as "breathing" or "wander"), which is attributable in part to backlash, is often encountered. This difference between the author's analysis and actual results may be a consequence of not including coulomb friction effects. Regardless, this difference does emphasize the need for continuing effort directed toward correlating analytical and experimental results.

For the low-amplitude oscillations generally resulting from

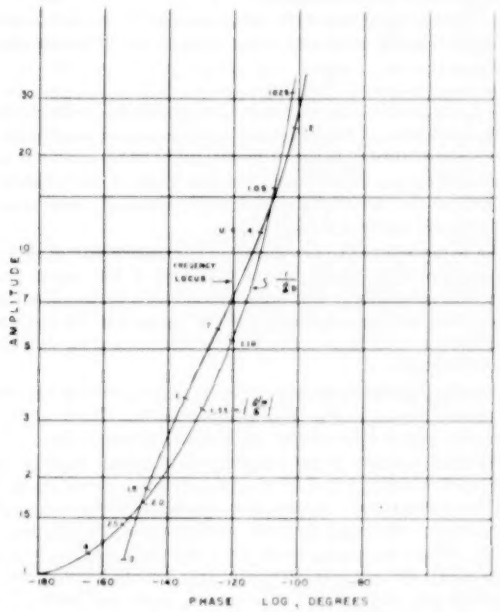


FIG. 36 LOG-ANGLE PLOT OF BACKLASH-DESCRIBING FUNCTION AND FREQUENCY LOCUS OF $\frac{5}{j\omega(j\omega + 1)}$

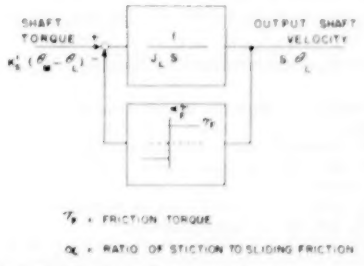


FIG. 37 COULOMB FRICTION REPRESENTATION

[6] "Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanism," by R. J. Kochenburger, Trans. AIEE, vol. 69, 1950, pp. 270–284.

backlash, the small amount of viscous damping has only negligible effect. Referring to the diagram in Fig. 11 of the paper, it would appear that a more exact representation would be to replace the viscous damping D_L with a combined friction term (see Fig. 37 of this discussion). With the representation of Fig. 37, many of the phenomena observed in actual practice can be duplicated on an analog computer. There is, however, apparently no method available for straightforward paper analysis of a system containing multiple nonlinear elements of this type, and as a result synthesis for such a system reduces to cut and try. Fortunately, in many systems, friction acts as a stabilizing factor and a system designed without consideration of friction will exhibit more stable characteristics than expected. This, however, is not always true for conditionally stable systems.

D. W. St. Clair.[7] In a linear system the order of the links in a single chain is immaterial; in a nonlinear system this is not necessarily true. Comment on the following lines by the author will be appreciated greatly:

1 Are there certain types of nonlinearities for which their location in an otherwise linear chain is still essentially immaterial? If so, what are examples?

2 Can any general rules be made relevant to the best location of an undesirable nonlinear component in an otherwise linear chain?

R. E. Kalman.[8] The author is to be commended for a stimulating treatment of the effects of saturation and dead band in closed-loop control systems. Understanding the behavior of nonlinear systems is a difficult problem and much of the progress in this field at the present time must come necessarily from analyzing particular types of nonlinearities.

The writer would like to suggest an alternative way of arriving at the qualitative conclusions of that part of the paper which deals with saturation. The approach is based on the now fairly well-known root-locus method[9,10] which shows how the poles of a closed-loop system move around in the complex s-plane as a function of loop gain.

A qualitative sketch of the root locus for the system in Fig. 16 of the paper is shown in Fig. 38 of this discussion. The direction of increasing gain is indicated by the arrows; the main point of interest is the behavior of the complex pair of poles. Suppose that these are located at A and A' when the system is operating at full gain (no saturation). As the gain decreases (due to saturation somewhere in the loop) the poles will move, say, to the point B and B'. Since the transients in the system are oscillatory (complex poles!), the instantaneous gain also will oscillate; therefore the poles may be thought of as moving back and forth on the segments A-B and A'-B'. Thus the system will be in the region of positive damping (left half-plane) some of the time and in the region of negative damping (right half-plane) the rest of the time. Heuristically, one would expect that when the average damping is zero (when negative damping is just canceled by positive damping during one period) the system will oscillate with steady amplitude.

If the root locus crosses over into the right half-plane only in a very small region as indicated by the dotted lines, it is plausible to assume that the negative damping never could be sufficiently large to offset the positive damping and therefore the oscillations would always decay. Thus the presence of right half-plane poles for some state of saturation is a necessary but probably not a sufficient condition for steady-state oscillations. Unfortunately, while the foregoing argument is quite picturesque and easy to interpret qualitatively, it does not appear to lend itself readily to quantitative treatment.

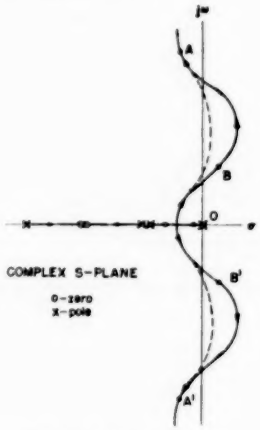


Fig. 38 Qualitative Root Locus of the Transfer Function of Author's Fig. 16

It is interesting to note that N. Levinson and O. K. Smith, in a now classical paper,[11] prove rigorously that stable oscillations will always exist if the poles of a second-order system move monotonically (crossing the $j\omega$-axis only once) from the right half-plane into the left half-plane as the amplitude of the signals in the loop is increasing. (The theorem has been rephrased to suit engineering terminology.) However, the writer is not aware of any mathematical theorem which would apply to high-order cases like the one discussed here.

Author's Closure

Mr. Gaines has pointed out two limitations of the approximate frequency-response method of analysis, namely, lack of accuracy and inability to handle multiple nonlinearities. As was indicated in the paper itself, the describing-function method is an approximate one and represents a means for arriving at results that are useful for engineering purposes although they may not be exact. The exact solutions for multiple loop systems having multiple nonlinearities are lengthy and practically prohibitive in the amount of work involved for hand solution. Fortunately, the use of analog computers as referred to by Gaines is becoming more widespread and makes possible the more exact study of complex systems with multiple nonlinearities.

Concerning the lack of accuracy of the describing-function method of analysis as illustrated by the second-order servo of Fig. 34 and the corresponding plot of Fig. 36, it will be noted that the sustained-oscillation conditions indicated by the describing function and frequency locus compare very favorably with the exact solution. The fact that the frequency-response method erroneously indicates stable operation for very low amplitudes of input disturbances is of little practical significance since input disturbances

[7] Control Engineer, Eastman Kodak Company, Rochester, N. Y. Mem. ASME.

[8] Research Assistant, Servomechanisms Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[9] "Control System Synthesis by Locus Methods," by W. R. Evans, Trans. AIEE, vol. 69, 1950, pp. 66–69.

[10] "The Study of Transients in Linear Feedback Systems by Conformal Mapping and the Root-Locus Method," by Victor C. M. Yeh, Trans. ASME, vol. 76, 1954, pp. 349–361.

[11] "A General Equation for Relaxation Oscillations," by N. Levinson and O. K. Smith, *Duke Mathematical Journal*, vol. 9, 1942, pp. 382–403.

of appreciable amplitude compared to the backlash will be encountered in any actual control system.

The condition of "breathing" mentioned by Mr. Gaines is one that is of considerable importance in position-control systems and was not revealed in the studies described in this paper. Extensions of these studies have indicated the importance of coulomb friction in producing this objectionable low-frequency cyclic motion. The simultaneous presence of dead-band and coulomb friction represents a condition of multiple nonlinearity that is not readily handled by the simple approximate methods. The presence of each of these nonlinearities separately, however, can be handled by the use of the frequency-response methods. Additional comparisons of analytical and analog-computer work is indicated to help find a way in which paper analyses of multiple nonlinearities may be performed.

The questions raised by Mr. St. Clair are indeed pertinent ones, but, unfortunately, the author is not in a position to provide generalized answers suitable for all conditions of nonlinearity. As illustrated in the body of the paper, the effect of a nonlinearity such as saturation is influenced by the remainder of the elements in the control. For example, if the nonlinearity occurs before the filter elements, Fig. 17, its effect may be quite different quantitatively than if the nonlinearity occurs after the filter elements, Fig. 18. It should be noted that a much greater amplitude of reference input may be required to saturate the power amplifier in Fig. 18 than is needed to saturate the preamplifier of Fig. 17. Experience seems to show that saturation-type nonlinearities are less objectionable if they follow the principal stabilizing networks and time constants rather than if they precede them.

Because of the wide variety of nonlinear "undesirable components" that may be encountered in a complex control system, it is impossible to generalize effectively as to the best location of the nonlinear element. As is indicated in the body of the paper, the designer can get an approximation to the effect of the given nonlinearity in a particular location by considering the equivalent sinusoidal response of the element to a sinusoidal input of the amplitude apt to be encountered in practice. This equivalent sinusoidal response may then be included in a linear analysis of the system to determine its effective stability and performance for this condition.

Mr. Kalman's physical interpretation of the phenomenon of saturation in terms of the root-locus method provides an alternative way of looking at this problem. The qualitative results of this approach are in general agreement with the results obtained by the exact analysis and the frequency-response methods. It is not clear that the approximate root-locus method affords any special advantage in terms of accuracy of quantitative results over the approximate frequency-response method described in the body of the paper. The question Kalman raises of whether the presence of right-half plane poles for some state of saturation is a necessary and/or sufficient condition for steady-state oscillation is an interesting one and might warrant additional investigation.

The comments of Messrs. Gaines, St. Clair, and Kalman serve to emphasize the importance of being able to handle nonlinearities in control-systems analysis and design. The methods outlined in this paper indicate ways in which the linear methods of frequency-response analysis may be utilized in obtaining approximate results for nonlinear systems.

Stability Characteristics of Closed-Loop Systems With Dead Band

By C. H. THOMAS,[1] SCHENECTADY, N. Y.

This paper presents a method of analyzing closed-loop systems under cyclic motion of the self-sustaining type, as caused by the presence of dead band. One common form of dead band considered occurs in mechanical linkages in conjunction with spring action, inertia, and viscous damping. The analysis is shown to be facilitated by means of an equivalent system representation. Simpler types of dead band, derivable from the more complex case, are considered. Stability criteria are developed, and the conditions are established under which sustained oscillations occur. The frequency and amplitude of oscillation may be predicted satisfactorily. Furthermore, the results are in a form which clearly indicates ways of suppressing the undesirable oscillations. One way of eliminating the hunting caused by dead band is shown. In support of the analysis the results of a differential-analyzer study are presented.

Nomenclature

The following nomenclature is used in the paper:

J_m = inertia of servomotor
J = inertia of controlled member
B_m = viscous damping constant of servomotor
B = viscous damping constant of controlled member
K = linear spring constant
b = dead band or backlash
r = constant relating motion between two points in mechanical linkage
K_v = velocity constant of position-control system
θ_m = position of servomotor
θ_c = position of controlled member
θ_r = reference position
ϵ = position error
α = input to dead-band element
A_m = maximum amplitude of α
β = output of dead-band element
μ_m = undamped natural frequency of servomotor
μ = undamped natural frequency of controlled member
T = time constant
ω = angular frequency, radians per sec
t = time, sec
s = complex variable of Laplace transform theory
$j = \sqrt{-1}$
e = 2.718 . . . . , base of natural logarithms

Introduction

THE functional similarity of the various types of automatic-control systems is well known. The versatile frequency-response method of analyzing such systems, first introduced in feedback-amplifier design, has been developed extensively in the servomechanism field.[2] In the development of the underlying theory, attention was first confined to systems which could be described by linear differential equations. Automatic control systems, whether they consist of electrical, mechanical, or hydraulic components, function in a linear fashion only over definite intervals of the variables. Frequently, analysis based on small excursions of the variables in the vicinity of a reference point allows specification of a system by linear differential equations with constant coefficients. This leads to great simplification, and is justified in part by the fact that many control systems operate substantially in this manner. If such analysis shows that small departures of the variables from all possible reference points vanish with time, stability under small oscillations is assured. The simplification of the problem in this manner has met with great success in the design of automatic controls and in extending the general theory of closed-loop systems. It is worth noting, however, that a system stable under small oscillations is not necessarily stable under large oscillations (5, 6).

In systems where linear small-oscillation theory is inapplicable, other methods of analysis must be used. Early approaches to strictly nonlinear control problems involved the direct solution of the nonlinear differential equations. If the calculations are carried out by hand, without the aid of mechanical or electronic computers, the procedure is usually very lengthy in all but the most simple cases. A new approach to this particular problem has evolved during the past few years. The frequency-response method of analysis and synthesis of linear systems has been extended to include ways of applying this same technique to a variety of essentially nonlinear problems.

Kochenburger (7) extended and successfully applied the sinusoidal method of analysis to relay servomechanisms; Tustin (8) used the same technique in studying one form of backlash; and Johnson (9) has demonstrated that more extensive application of the method is possible.

Dead zone, dead band, or backlash in an automatic control adversely affects the performance of the system, and in many instances is the direct cause of small-amplitude sustained oscillations. The analysis of closed-loop systems having dead band, as well as the analysis of systems having other types of nonlinearity, may be simplified by first deriving an equivalent system in which the linear and nonlinear elements are separated. The basic components of the system then consist (*a*) of linear elements which are frequency-variant, and (*b*) nonlinear elements which are approximately amplitude-variant. The concepts of frequency response then may be applied in a straightforward manner, and in many cases without undue complication.

Simple Types of Inactive Zone

The insensitive zones of measurement and response associated with various components of automatic-control systems not only

[1] Engineer, General Electric Company.

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, New York, N. Y., November 29–December 4, 1953, of The American Society of Mechanical Engineers.

[2] References (1–4) are representative of the literature. Numbers in parentheses refer to the Bibliography at the end of the paper.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 22, 1953. Paper No. 53—A-26.

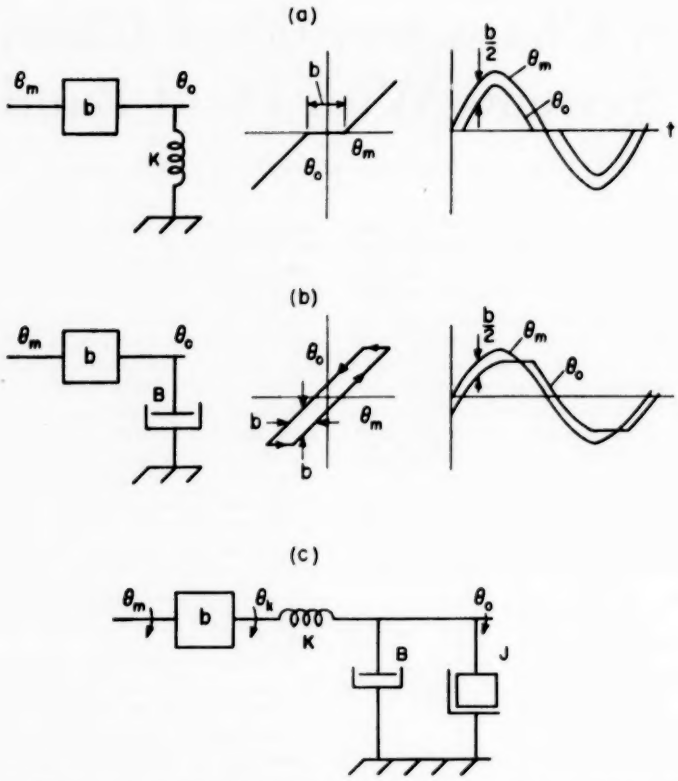


FIG. 1 MECHANICAL NETWORK DIAGRAMS REPRESENTING SIMPLE TYPES OF DEAD BAND

reduce accuracy but may impair the dynamic performance seriously. This loss in performance is not all attributable to time delay of initiating corrective action in the controller, as caused, for example, by dead zone in the error-detection mechanism, but unavoidably may be introduced to prevent or reduce sustained cycling of the controlled variables. Dead zones may occur, for example, in pressure-responsive elements as in gages and flow meters. Dead band or free-play may appear in mechanical linkages, as occasioned by nonprecision design or undue wear through usage. In servomechanisms backlash may appear in inferior gear trains.

A particular dead zone or dead band may be reduced to one of a few general types, depending on the immediately associated mechanical or electrical elements. Examples of some of the most simple types of elements having inactive zones are illustrated in Fig. 1.

In Fig. 1(*a*) a driving member transmits motion through a dead zone to a spring-loaded member. The transfer characteristic shown is valid if the output member has negligible inertia and friction.

In this example both the input and output variables are position. A similar characteristic may occur in a liquid-level or flow-control system, as caused by overlap of the pilot valve. In this instance the input variable is position of the pilot valve, and the output variable is velocity of the controlled piston.

In Fig. 1 (*b*) is depicted dead band and an output member characterized by friction alone. Such a case may occur where heavy friction loading obviates the effect of inertia on the output. With the input under oscillatory motion the backlash plus friction yields a hysteresis loop. Johnson (9) considered this type of nonlinearity and showed the transfer characteristic to be amplitude-variant only.

In Fig. 1(*c*) a controlled member characterized by inertia, or mass, and viscous damping B is driven through backlash and a linkage represented as a spring K. The transfer characteristic in this case cannot be so simply constructed as in the two former examples. An equivalent representation of the given mechanical network (to be discussed later), however, allows the behavior of the system to be more easily determined.

The sinusoidal-response characteristics of the foregoing dead-band elements can be described simply on specifying a reference point about which oscillation occurs. For servomechanisms with no unbalanced loading, as a result, for example, of constant torques acting on the output, the significant reference point is the origin of the co-ordinate axes, Fig. 1. For such cases the response to a sinusoidal input signal is a symmetrical oscillation consisting of fundamental and odd higher integer values of frequency.

The effect of dead band in closed-loop systems has been studied previously. In 1941 Concordia, Crary, and Parker (10) presented a comprehensive differential-analyzer study of the effect of dead band on the speed governor of power-system prime movers.

More recently Oldenbourg and Sartorius[3] showed how a dead band of the type shown in Fig. 1(*b*) could cause sustained oscillations in a closed-loop system.

Tustin (8) analyzed the case of dead band plus inertia. He also considered the effect of friction but neglected spring action. He utilized the frequency-response method of analysis and de-

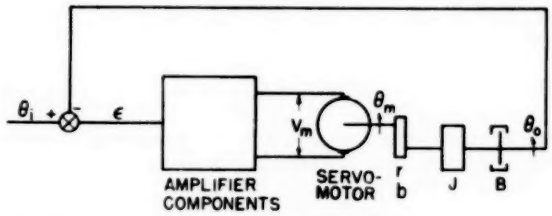
[3] Reference (11), pp. 169–174.

rived an approximating transfer function based upon the elements output response to a sinusoidal input signal.

As early as 1940, Charles Concordia applied the frequency-response concept in studying the effect of dead band on a speed governor. He did not develop or publish the method of analysis, however, since at that time his main interest was in determining power-system frequency variation and tie-line power swings caused by dead band in the prime-mover speed governor.

Description of System Studied

A single-loop servomechanism, shown in Fig. 2, is representative of the type of system studied in this paper. A servomotor



θ_i, reference position
θ_o, position of controlled member
ϵ, error
θ_m, position of servomotor
r, gear-train ratio
b, backlash in gear train
J_m, inertia of servomotor
J, inertia of controlled member
B, viscous-friction constant

Fig. 2 Block Diagram Showing Elements Associated With Backlash

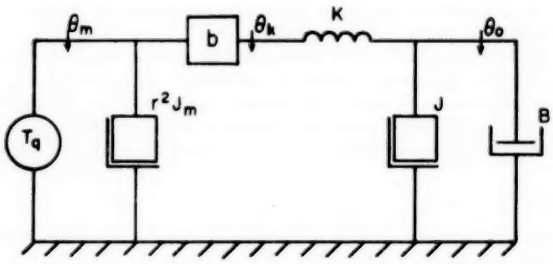


Fig. 3 Equivalent Mechanical Network Diagram

drives the controlled member through a gear train having backlash. The servomotor may be either a direct-current motor, as indicated in the figure, or a motor of the hydraulic or pneumatic type. The reader is referred to the nomenclature for a definition of symbols.

An equivalent mechanical network representing the servomotor and controlled member is shown in Fig. 3. There is a source of torque T_q. The backlash is denoted by b. An equivalent spring, representing the combined effect of all linear deformation of mechanical parts between motor and load, is assumed to be lumped on the load side of backlash. In the diagram all quantities are referred to the load side of the gear train.

In an actual case the distribution of inertia and spring action may require a more complex representation. Analysis, however, may proceed along the same general lines to that outlined.

Equivalent System

The transfer characteristics of the elementary forms of dead band, Figs. 1(*a*) and (*b*), are functions of signal amplitude and reference point only. The transfer characteristics of the system shown in Fig. 3 cannot be so simply described, since such characteristics are a function of frequency as well as signal amplitude and reference point. The problem can be reduced to simple terms, nevertheless, by first deriving an equivalent system which lends itself more readily to analysis.

Torque is transmitted through dead band b via spring K to a load represented by inertia J and viscous damping B, Fig. 3. Equating torques

$$K(\theta_k - \theta_o) = J\frac{d^2\theta_o}{dt^2} + B\frac{d\theta_o}{dt} \quad \ldots\ldots [1]$$

With the motor angle θ_m referred to the load side of the gear train and with backlash taken up

$$\theta_m - \theta_o = \theta_k - \theta_o \pm \frac{b}{2} \quad \ldots\ldots [2]$$

To assist in preparing a block diagram Equation [1] is Laplace transformed and a transfer function between θ_o and $(\theta_k - \theta_o)$ is formed. Thus

$$\frac{\theta_o(s)}{\theta_k(s) - \theta_o(s)} = \frac{\mu^2 T}{s(Ts+1)} \quad \ldots\ldots [3]$$

where $\mu^2 = K/J$, $T = J/B$, and $s = \sigma + j\omega$.

By Equation [3], a block is drawn for the linear components as in Fig. 4. The diagram is completed by representation of Equation [2]. The derived system is exact and of course valid for any form of input signal. The input to the nonlinear characteristic is $(\theta_m - \theta_o)$ and the output is $(\theta_k - \theta_o)$ where both variables are functions of time. For representation on a computer such as the differential analyzer, the required characteristic may be

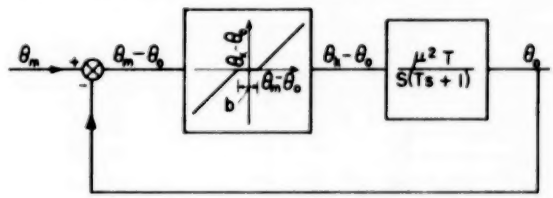


Fig. 4 Equivalent System

achieved by means of an automatic curve follower. The transfer function, represented by a single block, serves to describe the linear elements and is to be interpreted as a simplified representation of Equation [1].

The servomotor is assumed to be linear having speed-torque characteristics of negative slope B_m. With the variables expressed in per unit of chosen bases, the speed-torque characteristics are defined by

$$T_q = V_m - \frac{d\theta_m}{dt} \quad \ldots\ldots [4]$$

where V_m is the input signal to the servomotor, T_q is the servomotor torque, and $(d\theta_m)/(dt)$ is the servomotor velocity. The torque developed by the servomotor accelerates its own mass and transmits torque to the load when the dead band is taken up. The equation of motion may be written in the form

$$T_m\frac{d^2\theta_m}{dt^2} + \frac{d\theta_m}{dt} + T_m\mu_m^2(\theta_k - \theta_o) = V_m \quad \ldots\ldots [5]$$

The servomotor time constant is $T_m = J_m/B_m$. The constant μ_m has the dimensions of radian frequency and is defined by

$$\mu_m = \sqrt{\frac{K}{r^2 J_m}}$$

Laplace transforming Equation [5] and solving for motor position

$$\theta_m(s) = \frac{V_m(s) - T_m \mu_m^2 [\theta_k(s) - \theta_o(s)]}{s(T_m s + 1)} \quad [6]$$

The block diagram for the servomotor including backlash and load elements may now be constructed. This is shown in Fig. 5. The remaining part of the servomechanism has the transfer function $Q(s)$. The loop is closed by feeding back the output angle θ_o.

Stability Considerations

All components of the system shown in Fig. 5 are linear except for the element representing dead band. If the input to the dead-band element is Laplace transformable, then

$$A(s) = \theta_m(s) - \theta_o(s) \quad [7]$$

Similarly, the output is

$$B(s) = \theta_k(s) - \theta_o(s) \quad [8]$$

The system is viewed from this point regarding $B(s)$ as the input to the system and $A(s)$ as the output. A Laplace transformed function relating $A(s)$ to $B(s)$ around the linear part of the system may be derived without approximations. Thus from the figure

$$\theta_o(s) = G_1(s) B(s) \quad [9]$$

$$A(s) = \theta_m(s) - \theta_o(s) \quad [10]$$

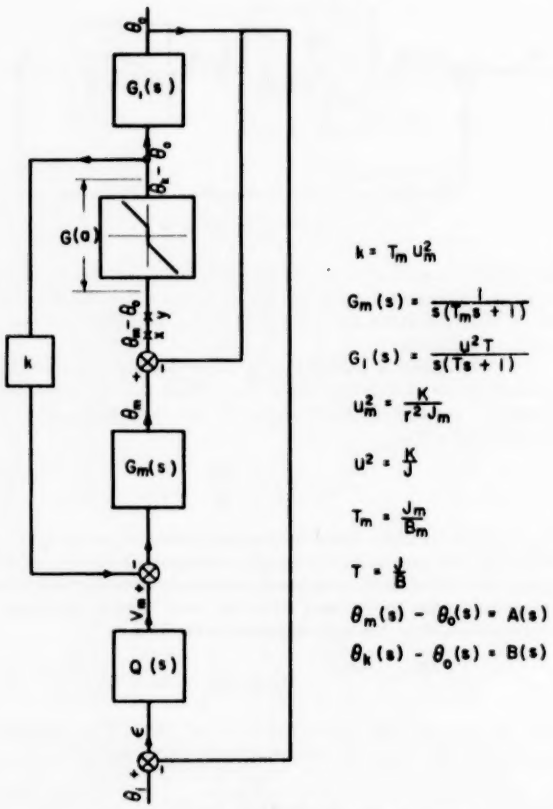


Fig. 5 Block Diagram of System Studied

$$\theta_m(s) = G_m(s) [V_m(s) - kB(s)] \quad [11]$$

and

$$V_m(s) = Q(s) [\theta_i(s) - \theta_o(s)] \quad [12]$$

The linear transfer functions are defined by

$$G_1(s) = \frac{\mu^2 T}{s(Ts + 1)} \quad [13]$$

$$G_m(s) = \frac{1}{s(T_m s + 1)} \quad [14]$$

$$k = T_m \mu_m^2 \quad [15]$$

and

$$Q(s) = \frac{V_m(s)}{E(s)} \quad [16]$$

For a stability study all inputs to the system from external sources are zero; hence Equation [12] becomes

$$V_m(s) = -Q(s) \theta_o(s) \quad [17]$$

From the foregoing set of equations

$$G_1(s) + [k + Q(s) G_1(s)] G_m(s) = -\frac{1}{\frac{B(s)}{A(s)}} \quad [18]$$

The linear transfer functions appear on the left-hand side of Equation [18] and are equated to the negative reciprocal of the ratio of dead-band output to input signals.

The next step is to test Equation [18] for equality under the condition of constant-amplitude oscillations. If sustained oscillations are occurring in the system, then both $B(s)$ and $A(s)$ are functions representable by Fourier series in the time domain. Suppose the input function $\alpha(t)$ to the dead band is of the form

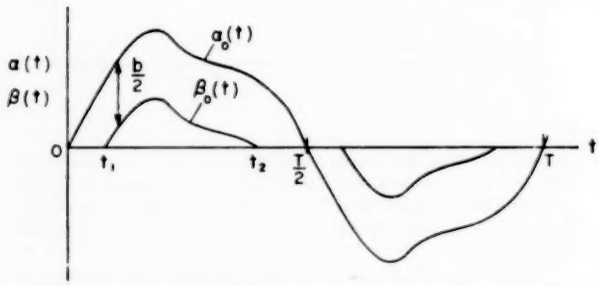


Fig. 6 Relationship of Dead-Band-Input and Output Oscillations

shown in Fig. 6 where $\alpha_n(t)$ defines the oscillation over one period T such that

$$\alpha_n(t - nT) = 0, \text{ outside interval } nT < t < (n+1)T \quad [19]$$

The oscillation then is expressed by the sum

$$\alpha(t) = \sum_{n=0}^{\infty} \alpha_n(t - nT)\, u(t - nT) \quad [20]$$

where $u(t - nT)$ is the unit step function shifted to the right by nT. Applying the real translation theorem, the Laplace transform of Equation [20] is

$$A(s) = A_0(s) \sum_{n=0}^{\infty} e^{-nTs} \quad [21]$$

where by definition

$$A_0(s) = \int_0^T \alpha_0(t)\, e^{-st}\, dt \quad [22]$$

The infinite geometric series in Equation [21] may be summed yielding

$$A(s) = A_0(s) \frac{1}{1 - e^{-Ts}} \quad [23]$$

Similarly, the Laplace transformed output of the dead-band element is

$$B(s) = B_0(s) \frac{1}{1 - e^{-Ts}} \quad [24]$$

where

$$B_0(s) = \int_0^T \beta_0(t)\, e^{-st}\, dt \quad [25]$$

$\beta_0(t) = 0$, outside interval $0 < t < T$. From Equations [23] and [24]

$$\frac{B(s)}{A(s)} = \frac{B_0(s)}{A_0(s)} \quad [26]$$

Thus the ratio of the Laplace transforms of the repeated oscillations is equal to the ratio of transforms over one period.

Substituting Equation [26] in Equation [18]

$$H(s) = -\frac{1}{\dfrac{B_0(s)}{A_0(s)}} \quad [27]$$

where for brevity the transfer function around the linear part of the closed loop is

$$H(s) = G_1(s) + [k + Q(s)\, G_1(s)]\, G_m(s) \quad [28]$$

If sustained oscillations of frequencies $n\omega$ are occurring, it is clear that Equation [27] must be satisfied for each harmonic frequency separately. Thus setting $s = jn\omega$

$$H(jn\omega) = -\frac{1}{\dfrac{B_0(jn\omega)}{A_0(jn\omega)}} \quad [29]$$

For a given frequency the relative amplitude and phase shift through the linear part of the system must be equal to the negative reciprocal of the relative amplitude and phase across the dead-band element. This conclusion is self-evident and follows immediately on assuming the existence of sustained oscillations.

Another way of viewing the foregoing is outlined later. Suppose the input to the nonlinear element, Fig. 5, is broken between the points x and y. Presume a hypothetical signal generator capable of generating any frequency and all integer multiples of frequency is connected at y. If the relative amplitudes and phases of all harmonic terms can be adjusted so that the signal is exactly reproduced at x, then the transmission ratio around the system is unity for all frequencies. The frequencies and amplitudes of sustained oscillation are thus determined, and on connecting x to y and removing the hypothetical signal generator the oscillations continue indefinitely.

The foregoing is based on the hypothesis that sustained oscillations can develop in the closed-loop system. Proof is offered in the fact that such oscillations are observed to occur in physical systems of this type. Furthermore, as will be shown, an analog-computer representation of the system in question develops oscillations of the type described.

Transfer Characteristics of Dead-Band Element

The explicit determination of the dead-band transfer characteristics

$$\frac{B_0(jn\omega)}{A_0(jn\omega)}$$

depends upon knowledge of the form of the input oscillations. If the form of the input $\alpha(t)$ is known, then of course the output $\beta(t)$ of the nonlinear element is determined easily. The analysis would be too complex for general use if it were required to derive a new set of transfer ratios for each individual case. Fortunately, such a procedure is unnecessary. Oscillations obtained on the differential analyzer for a wide variety of conditions show that the input $\alpha(t)$ to the nonlinear element may be approximated satisfactorily by a single sinusoid.

With all possible values of the fundamental frequency-transfer ratio known, the conditions for sustained oscillation are ascertained easily. In so doing the dependency of the oscillation amplitude and frequency on certain controlled variables (such as the velocity constant) in the system may be established. If the analysis is carried to the point of accounting for the higher frequencies, the principal objective is to determine their effect on the fundamental frequency-transfer ratio.

With the input to the nonlinear element approximating a sinusoid, the output will have the form shown in Fig. 7. Oscilla-

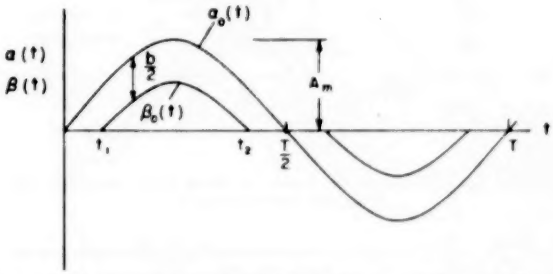


Fig. 7 Relationship of Dead-Band Input and Output Wave Forms for Sinusoidal Input

tion about the origin is assumed, the case of a servomechanism having no unbalanced loading. The solution of Equation [26] for this case appears in the Appendix. The result obtained is

$$G(\alpha) = 1 - \frac{2}{\pi}\left[\sin^{-1}\frac{b}{2A_m} + \frac{b}{2A_m}\sqrt{1 - \left(\frac{b}{2A_m}\right)^2}\right] \quad [30]$$

where

$$G(\alpha) = \frac{B_0(j\omega)}{A_0(j\omega)}$$

the fundamental frequency-transfer ratio.

The transfer ratio is a real positive number dependent only on the ratio of dead band to twice the peak amplitude of the input wave.

This expression shows that $G(\alpha)$ can have values only in the range

$$0 \leq G(\alpha) \leq 1.0$$

If the amplitude of the input wave is such that $2A_m \leq b$, then

$G(\alpha) = 0$. But as $A_m \rightarrow \infty$, $G(\alpha) \rightarrow 1.0$. Thus as $G(\alpha)$ increases the amplitude of the associated oscillation increases. In Fig. 8 is plotted $G(\alpha)$ versus $b/2A_m$.

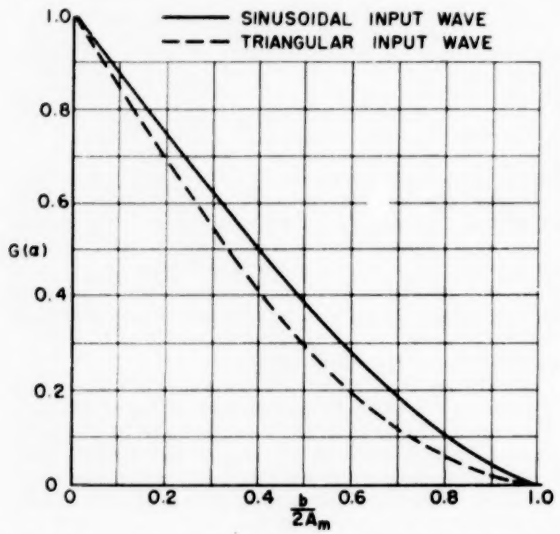


FIG. 8 COMPARISON OF DEAD-BAND TRANSFER RATIOS FOR TWO TYPES OF INPUT WAVES

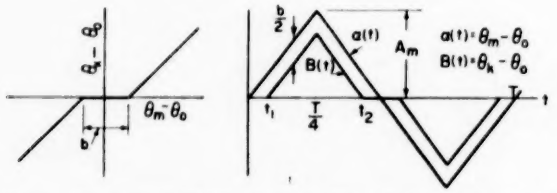


FIG. 9 TRIANGULAR WAVE INPUT TO DEAD-BAND ELEMENT AND CORRESPONDING OUTPUT

If there is little or no friction associated with the output member, the oscillation $(\theta_m - \theta_o)$ approaches a triangular wave. In such cases the relationship of the input $(\theta_m - \theta_o)$ to the output $(\theta_k - \theta_o)$ may be approximated as shown in Fig. 9. The evaluation of Equation [26] (see Appendix) yields

$$\frac{B_0(jn\omega)}{A_0(jn\omega)} = 1 - \frac{\sin \frac{n\pi}{2} \frac{b}{2A_m}}{\sin n\frac{\pi}{2}} \quad \ldots\ldots\ldots [31]$$

where n is an odd integer. Setting $n = 1$ the fundamental frequency-transfer ratio is

$$\frac{B_0(j\omega)}{A_0(j\omega)} = 1 - \sin \frac{b}{2A_m} \frac{\pi}{2} \quad \ldots\ldots\ldots [32]$$

A comparison is made in Fig. 8 between the triangular wave-transfer ratio and the transfer ratio obtained assuming a sinusoidal input. In both cases there is no phase shift between fundamental components of frequency across the dead-band element. This result is obtained of course from the symmetry of the input waves.

If the value of $G(\alpha)$ for which sustained oscillations would occur is computed for a particular case, $b/2A_m$ as obtained from Fig. 8 for a triangular wave would be somewhat less than the value obtained for a sinusoidal wave. For a given dead band then, the triangular wave results in larger amplitudes of oscillation. Since the phase of $G(\alpha)$ in the two cases is zero, there would be no difference in the two predicted fundamental frequencies of oscillation.

If lack of symmetry of the input wave results in appreciable phase shift across the dead-band element, the approximation in the foregoing analysis naturally becomes greater. A new dead-band amplitude variant transfer ratio may be derived, similar to the preceding, to provide greater accuracy in studying such cases. In the systems studied by the author, this was not found to be necessary.

As just demonstrated, and in the Appendix, the dead-band amplitude variant transfer function can be computed in a relatively easy manner by approximating the form of the input wave with a simple Laplace transformable function.

METHOD OF PREDICTING AMPLITUDE AND FREQUENCY OF OSCILLATION

If all possible dead-band transfer ratios

$$\frac{B_0(j\omega)}{A_0(j\omega)}$$

for the fundamental frequency are known, determination of system stability becomes a straightforward procedure. Test of Equation [29] for equality to determine if sustained oscillations are possible is equivalent to testing the open-loop transfer function

$$\frac{\theta_o(j\omega)}{E(j\omega)}$$

for intersection of the -1 point in the complex plane.

The open-loop transfer function for the system in Fig. 5 is derived from Equations [9] through [12]. The transfer ratio for the nonlinear element is replaced by $G(\alpha)$ which is defined in Fig. 8 for two separate forms of oscillation. Thus

$$\frac{\theta_o(j\omega)}{E(j\omega)} = Q(j\omega)\, G_1(j\omega)\, G_m(j\omega)$$

$$\frac{G(\alpha)}{1 + G(\alpha)\,[G_1(j\omega) + kG_m(j\omega)]} \quad [33]$$

The procedure then is to determine for any value of $G(\alpha)$ in its range of possible values, $0 \leqq G(\alpha) \leqq 1.0$, if the open-loop transfer function intersects the -1 point.

The simplest method of determining all possible -1 points of the open-loop transfer function, Equation [33], is a graphical one. To begin the computations, a frequency ω_1 is assumed, and the amplitude and phase

$$K_1/\phi_1 = Q(j\omega_1)\, G_1(j\omega_1)\, G_m(j\omega_1) \quad \ldots\ldots [34]$$

is determined. Next, the amplitude and phase

$$K_2/\phi_2 = G_1(j\omega_1) + kG_m(j\omega_1) \quad \ldots\ldots [35]$$

is computed. If the transfer-function locus intersects -1 at this frequency, then the net phase lag must be -180 deg. Thus

$$\text{Angle}\,[1 + G(\alpha)\, K_2/\phi_2] = \phi_3 = 180 + \phi_1 \quad \ldots\ldots [36]$$

At this point a line o-b is drawn at an angle ϕ_3 with the reference axis as illustrated in Fig. 10. Next, a line a-c is drawn at an angle ϕ_2 starting at $+1$ on the real axis. The intersection point p of the two lines drawn is then located. The value of $G(\alpha)$ fulfilling the phase-shift requirement then is

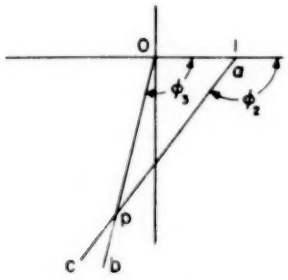


FIG. 10 GRAPHICAL METHOD OF DETERMINING $G(\alpha)$

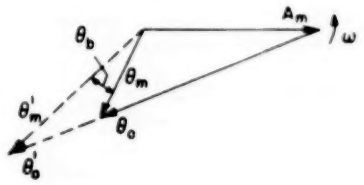


FIG. 11 PHASE LAG CAUSED BY DEAD BAND

$$G(\alpha) = \frac{ap}{K_2} \quad \ldots\ldots [37]$$

The net gain at the frequency ω_1 is

$$\left|\frac{\theta_o(j\omega_1)}{E(j\omega_1)}\right| = K_1 \frac{G(\alpha)}{op} = \frac{K_1}{K_2}\frac{ap}{op} \quad \ldots\ldots [38]$$

If the transfer-function locus is to pass through the -1 point, the system gain must be multiplied by a correction factor. From Equation [38]

$$\text{Gain correction factor} = \frac{K_2}{K_1}\frac{op}{ap} \quad \ldots\ldots [39]$$

From the foregoing the system gain which would produce constant-amplitude oscillations at the frequency ω_1 is determined. Repeating the foregoing for several values of frequency yields sufficient data to establish all possible frequencies and amplitudes of oscillation.

With the system gain or velocity constant known and the corresponding value of $G(\alpha)$ and frequency of oscillation determined, the fundamental frequency amplitude of oscillation at any point in the system can be easily determined. From Fig. 5

$$\theta_o(j\omega) = \frac{\mu^2 T}{j\omega(j\omega T + 1)} B(j\omega) \quad \ldots\ldots [40]$$

$$A(j\omega) = \theta_m(j\omega) - \theta_o(j\omega) \quad \ldots\ldots [41]$$

$$B(j\omega) = G(\alpha) A(j\omega) \quad \ldots\ldots [42]$$

For convenience the relative phase of $A(j\omega)$ will be taken as zero so that on a conventional vector diagram

$$A(j\omega) = A_m/0° \quad \ldots\ldots [43]$$

From the relations given the amplitude and phase of the load angle is

$$\theta_o(j\omega) = \frac{G(\alpha)}{j\omega(j\omega T + 1)} \mu^2 T A_m \quad \ldots\ldots [44]$$

The amplitude and phase of the servomotor angle (referred to load side of gear train)

$$\theta_m(j\omega) = A_m + \theta(j\omega) \quad \ldots\ldots [45]$$

A typical representation of the various amplitudes is shown in Fig. 11.

By definition, Equation [41], A_m represents the maximum displacement between motor and load positions. For the given value of $G(\alpha)$, the ratio $b/2A_m$ is obtained from the graph in Fig. 8. Thus A_m may be expressed in terms of the dead band b. Since A_m is directly proportional to b, the foregoing equations show that all amplitudes of oscillation are directly proportional to the dead band. Results from the differential analyzer confirm this fact.

If oscillations of amplitude A_m in Fig. 11 are forced on the system with no dead band present [$G(\alpha) = 1.0$], then the load and motor angles change to θ_o and θ_m, respectively. The angle by which the load lags the motor is decreased by θ_b. Dead band therefore introduces a phase lag into the system. This conclusion is easily reached from Equations [44] and [45] considering the fact that in the presence of dead band $G(\alpha)$ is always less than unity.

As the amplitude of oscillation approaches, the limiting value of transmission, $A_m \to b/2$, $G(\alpha)$ becomes negligible. The transfer function between output $\theta_o(j\omega)$ and error $E(j\omega)$, Equation (33), becomes approximately

$$\frac{\theta_o(j\omega)}{E(j\omega)} = Q(j\omega) G_1(j\omega) G_m(j\omega) G(\alpha)$$

$$= Q(j\omega) \frac{\mu^2 T G(\alpha)}{(j\omega)^2 (j\omega T + 1)(j\omega T_m + 1)} \quad \ldots\ldots [46]$$

This expression shows clearly the large phase lag (approaching -90 deg) caused by dead band at small amplitudes of oscillation.

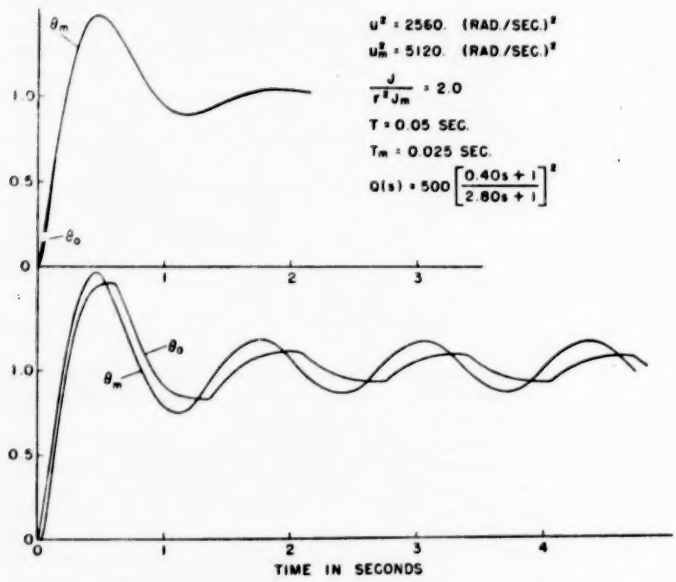
As the amplitude of oscillation decreases, the phase lag increases, invariably resulting in an intersection of the -1 point in the complex plane by the open-loop transfer-function locus.

DIFFERENTIAL-ANALYZER RESULTS

The closed-loop system in Fig. 5 was set up on the mechanical differential analyzer of the author's company. Typical responses obtained on the analyzer for a step change in reference signal are shown in Figs. 12 through 18. In each case the system response with and without dead band was obtained. In all cases oscillations developed when dead band was present.

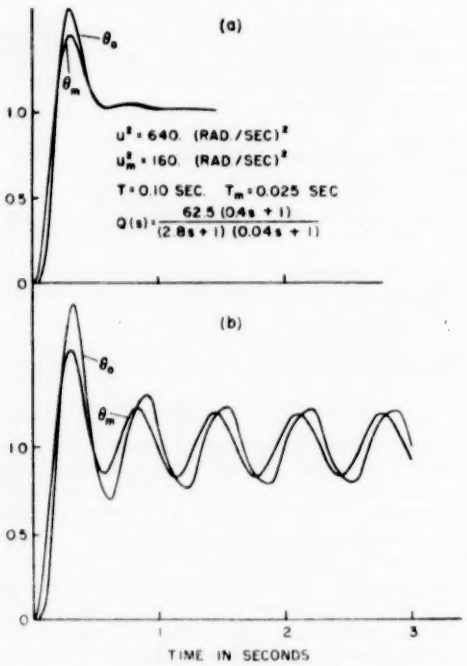
All analyzer runs were for unconditionally stable systems with the exception of the run shown in Fig. 12. The system in this instance was conditionally stable; that is, in order to be stable the open-loop gain K_v was required to be above a certain minimum value. Lowering of the gain in systems of this type results in instability.

The amplitudes of all responses are dimensionless, and the dead band b is expressed in per unit. Dead band is usually very small, and for accurate representation on a computing device the ratio of signal amplitude to dead band must not be great; otherwise, the magnitude of various quantities appearing as integrands would exceed the practical limits of the integrators. To illustrate, suppose one per unit in Fig. 12 represents 10 milliradians. The total dead band then is 2 milliradians. If one per unit is 1 deg, then the dead band is 0.2 deg.



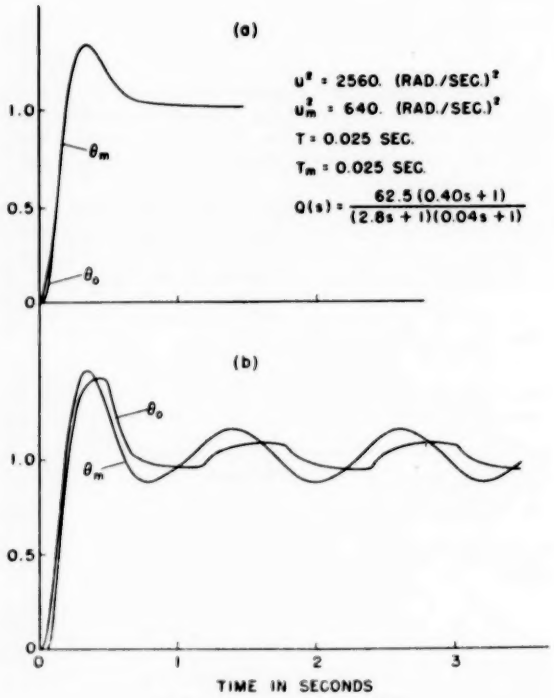
(*a*) $b = 0$ (*b*) $b = 0.20$

FIG. 12 RESPONSE TO UNIT STEP INPUT, SYSTEM CONDITIONALLY STABLE



(*a*) $b = 0$ (*b*) $b = 0.20$

FIG. 13 RESPONSE TO UNIT STEP INPUT



(*a*) $b = 0$ (*b*) $b = 0.20$

FIG. 14 RESPONSE TO UNIT STEP INPUT

CHECK OF STABILITY CRITERIA

For all the differential-analyzer responses shown, an analysis was made, using the method described earlier, for the purpose of predicting the amplitude and frequency of oscillation in each case. The results of these computations, showing predicted amplitudes and frequencies of oscillation in comparison with the actual values obtained on the differential analyzer, are given in Table 1.

The order of magnitude of all quantities was very well predicted. The analysis of course can be refined to account for higher harmonic frequencies present in the oscillations, but in servomechanism design such refinement is not at all necessary.

The significant point here is that stability criteria derived for a common type of backlash or dead band has been found valid and

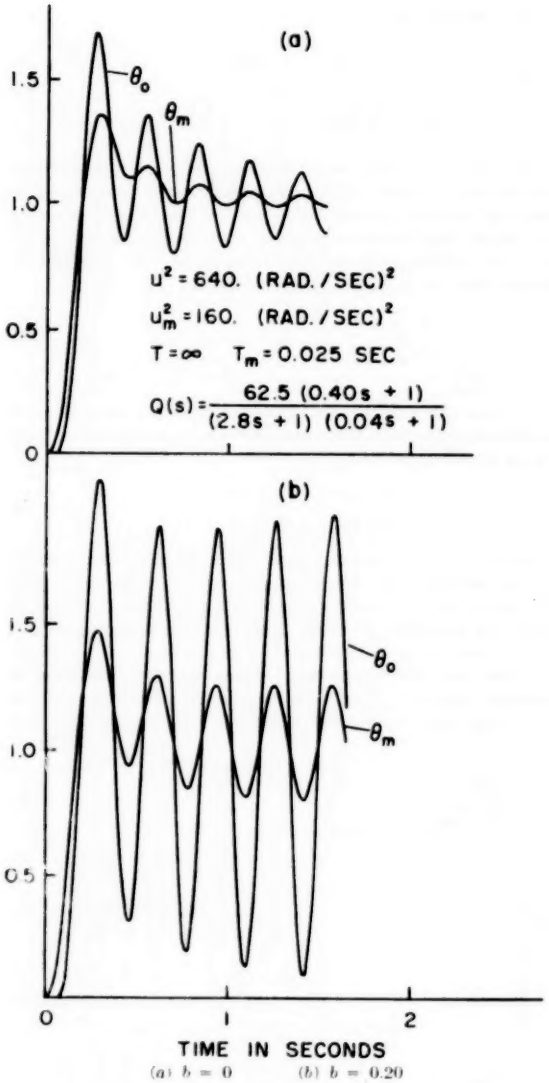


FIG. 15 RESPONSE TO UNIT STEP INPUT

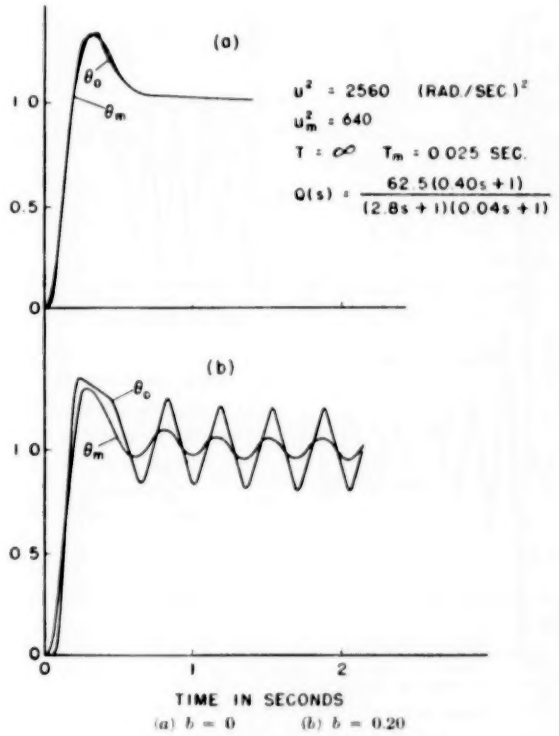


FIG. 16 RESPONSE TO UNIT STEP INPUT

therefore should be useful in predicting the stability of other closed-loop systems in which it occurs.

The synthesis of stabilizing components becomes possible. To demonstrate this, at least one method will be shown which eliminates altogether the oscillations caused by backlash.

SUPPRESSING OSCILLATIONS CAUSED BY BACKLASH

In seeking a means for suppressing or eliminating oscillations caused by backlash the designer might attempt incorporating linear cascade compensation. By appropriate design the amplitude of oscillation may be minimized. It would be extremely difficult, if not impossible, to suppress oscillations entirely by such a procedure. The locus of the open-loop function always undergoes a phase shift greater than —180 deg as the frequency approaches a high value, resulting in intersection of the negative real axis. A value of $G(\alpha)$, between zero and unity, always can be found which results in the locus intersecting the —1 point, the condition for sustained oscillations.

The foregoing conclusion is supported by the fact that no physically realizable component such as $Q(s)$, Fig. 5, can have a transfer function of order less than zero (the order of the transfer function is defined as the order of $j\omega$ in the denominator minus the order of $j\omega$ in the numerator). For the minimum possible order of $Q(s)$ sustained oscillation occurs.

Changing the system velocity constant cannot eliminate the oscillations caused by backlash. This fact is illustrated in Figs. 19 and 20. The velocity constant is defined as the output velocity per unit error input. Thus

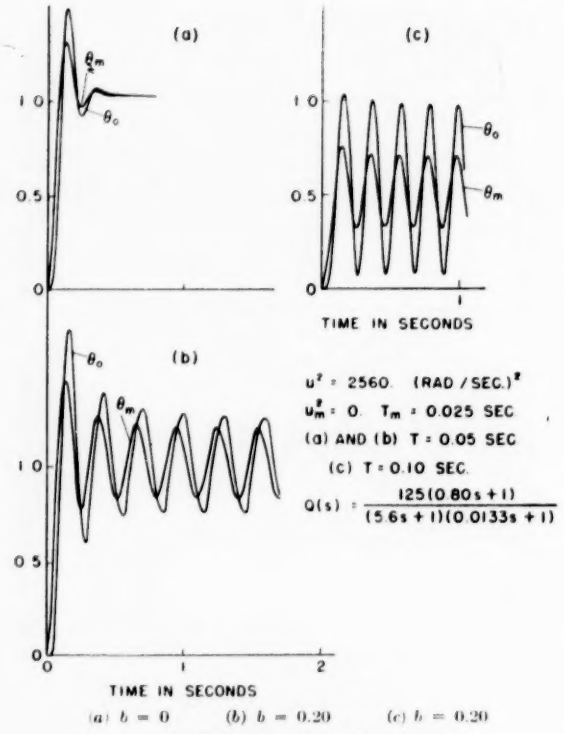
$$\lim_{s \to 0} s \frac{\theta_o(s)}{E(s)} = K_v$$

The curves shown in Figs. 19 and 20 were computed for the systems described in Fig. 17(*b*) and Fig. 16, respectively. In each case the frequency of oscillation and the ratio $b/2\theta_o$ is plotted as a function of system velocity constant. These curves show that reducing the velocity constant decreases both the frequency and

TABLE 1 COMPARISON OF PREDICTED AMPLITUDE AND FREQUENCY OF OSCILLATION WITH VALUES OBTAINED ON DIFFERENTIAL ANALYZER

		—Predicted values—			—Measured values—		
Fig. no.	$G(\alpha)$	Cycles per sec	$2\theta_o$ / b	$2\theta_m$ / b	Cycles per sec	$2\theta_o$ / b	$2\theta_m$ / b
12.....	0.025	0.81	0.65	1.12	0.77	0.82	1.54
13.....	0.330	1.62	2.64	1.86	1.50	3.14	1.88
14.....	0.070	0.89	0.94	1.38	0.82	0.74	1.43
15.....	0.780	3.10	7.52	1.80	3.16	9.54	2.44
16.....	0.20	3.10	1.96	0.52	2.90	2.02	0.50
17(*b*)...	0.40	3.58	3.36	2.14	3.33	2.60	1.88
17(*c*)...	0.58	4.87	4.60	1.98	4.84	4.58	1.88
18(*b*)...	0.49	4.54	2.40	0.44	3.94	2.20	0.44
18(*c*)...	0.49	4.54	3.76	1.24	4.54	3.58	1.18
18(*d*)...	0.49	4.54	2.40	0.44	4.17	2.30	0.44

NOTE: From the computed value of $G(\alpha)$, the value of $b/2A_m$ was obtained from Fig. 8. $2\theta_o/b$ and $2\theta_m/b$ were calculated by use of Equations [44] and [45], respectively.



(a) $b = 0$ (b) $b = 0.20$ (c) $b = 0.20$

FIG. 17 RESPONSE TO UNIT STEP INPUT

amplitude of sustained oscillation. The ratio $b/2\theta_o$ increases in value as K_v is decreased, but for constant backlash b, θ_o of course decreases.

Any device which eliminates the instability caused by backlash must counteract the large phase lag which occurs at small amplitudes of oscillation (see Equation [46]). One possible means of stabilization is shown in Fig. 21. The backlash and associated spring inertia, and viscous damping are represented as before. The added element is a viscous damper B_e connected between the servomotor and load shafts. This simple scheme is chosen merely to demonstrate the method of analysis. Other ways of suppressing oscillations, which are best suited for a particular case, may be determined by following similar procedure to that outlined.

The transfer function for the system shown in Fig. 21 may be derived by equating torques. Thus

$$J\frac{d^2\theta_o}{dt^2} + B\frac{d\theta_o}{dt} = K(\theta_b - \theta_o) + B_e\left(\frac{d\theta_m}{dt} - \frac{d\theta_o}{dt}\right) \quad [47]$$

For simplicity, a gear ratio of unity is assumed. For sinusoidal variations, Equation [47] becomes

$$j\omega(j\omega T + 1)\theta_o(j\omega) = T\mu^2[\theta_b(j\omega - \theta_o(j\omega)] + \frac{T}{T_e}j\omega[\theta_m(j\omega) - \theta_o(j\omega)] \quad [48]$$

where

$$T = \frac{J}{B}, \quad \mu^2 = \frac{K}{J}, \quad \text{and} \quad T_e = \frac{J}{B_e}$$

Use is made of the dead-band transfer ratio. Thus

$$\theta_b(j\omega) - \theta_o(j\omega) = G(\alpha)[\theta_m(j\omega) - \theta_o(j\omega)] \quad [49]$$

From the foregoing

$$\frac{\theta_o(j\omega)}{\theta_m(j\omega)} = \frac{j\omega\frac{T}{T_e} + T\mu^2G(\alpha)}{j\omega\left(1+\frac{T}{T_e}\right)\left(j\omega\frac{TT_e}{T+T_e}+1\right) + T\mu^2G(\alpha)} \quad [50]$$

For $G(\alpha) = 1.0$, the transfer function just given represents the case of zero backlash. If large-amplitude oscillations are occurring with backlash present $G(\alpha) \to 1.0$ (see Fig. 8) and the amplitude-phase characteristics in the two instances approach equality. For oscillations of very small amplitude $G(\alpha) \to 0$, and the transfer function becomes approximately

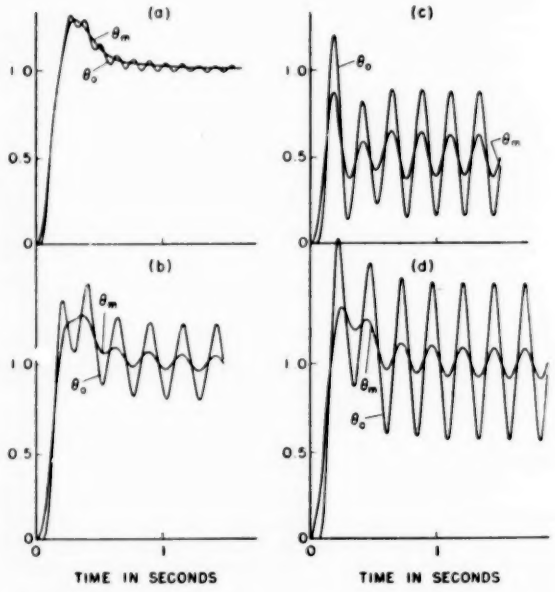
$$\frac{\theta_o(j\omega)}{\theta_m(j\omega)} = \frac{T}{T+T_e}\frac{1}{j\omega\frac{TT_e}{T+T_e}+1} \quad [51]$$

For very small amplitudes of oscillation in the usual range of possible unstable frequencies with the viscous damper B_e omitted, the applicable transfer function is

$$\frac{\theta_o(j\omega)}{\theta_m(j\omega)} = \frac{T\mu^2G(\alpha)}{j\omega(j\omega T+1)} \quad [52]$$

The effect of the viscous damper is readily seen. Without the viscous damper, an additional phase lag of —90 deg occurs. The viscous damper in effect removes the $1/(j\omega)$ term in Equation [52], thus providing the large positive phase shift required for stabilization at small amplitudes of oscillation.

In this example the problem becomes one of determining the required constant T_e for the viscous damper. The system chosen for this purpose is described in Fig. 13. The required value of



(a) $b = 0$ (b) $b = 0.2$ (c) $b = 0.2$ (d) $b = 0.4$

FIG. 18 RESPONSE TO UNIT STEP INPUT

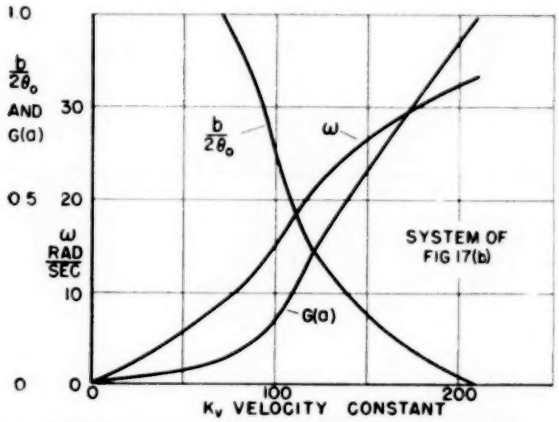


FIG. 19 DEPENDENCY OF FREQUENCY AND AMPLITUDE OF OSCILLATION ON THE SYSTEM VELOCITY CONSTANT

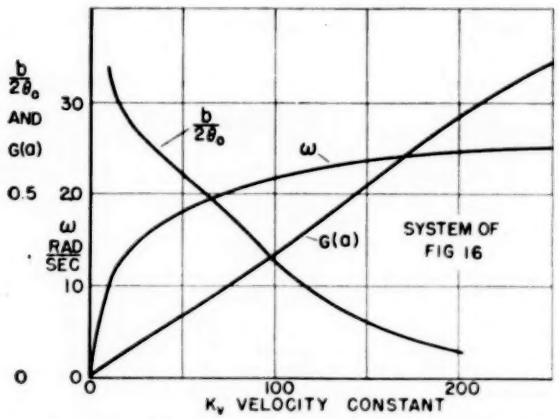


FIG. 20 DEPENDENCY OF FREQUENCY AND AMPLITUDE OF OSCILLATION ON THE SYSTEM VELOCITY CONSTANT

T_v was determined experimentally on an analog computer. With compensation sufficient to eliminate sustained oscillations, the output response of the system to a step change in reference signal was recorded. The wave forms obtained are shown in Fig. 22. With backlash present, sustained oscillations occur, Fig. 13(*b*), but with a viscous damper paralleling the backlash element the oscillations are damped out.

The design procedure suggested by the foregoing may be carried out in two steps: (*a*) Assume backlash to be zero and design the closed-loop system for acceptable performance. This eliminates the possibility of large-amplitude oscillations developing with backlash. The amplitude of any oscillation which does occur will be of the same order of magnitude as the backlash. (*b*) Account for backlash using the applicable transfer function and seek practical means of offsetting the large negative phase shift which occurs for small oscillations. At the appropriate stage in the design the open-loop transfer function should indicate closed-loop stability for extremely small values of the backlash transfer ratio $G(\alpha)$

LIMIT OF STABILITY—BACKLASH, SPRING, AND INERTIA

Important but simpler cases of dead band are derivable from the foregoing. For example, if the torque transmitted by the the gear train is negligible, under conditions of oscillation, in comparison with the servomotor's own inertial and friction torques, then

$$\frac{\mu_m^2}{\mu^2} = \frac{J}{r^2 J_m} \to 0$$

This condition is approached in small servomechanisms where large step-down gear trains are used.

With this approximation, the open-loop transfer function, Equation [33], becomes

$$\frac{\theta_o(j\omega)}{E(j\omega)} = Q(j\omega)\frac{1}{j\omega(j\omega T_m + 1)}\frac{G(\alpha)}{\frac{j\omega(j\omega T + 1)}{\mu^2 T} + G(\alpha)} \quad [53]$$

Examples of differential-analyzer runs for this case are given in Fig. 17. If viscous damping on the output shaft is negligible, $T \to \infty$, the transfer function reduces to

$$\frac{\theta_o(j\omega)}{E(j\omega)} = Q(j\omega)\frac{1}{j\omega(j\omega T_m + 1)}\frac{G(\alpha)}{-\left(\frac{\omega}{\mu}\right)^2 + G(\alpha)} \quad [54]$$

Differential-analyzer runs obtained for this case are shown in Fig. 18. With no backlash present, Fig. 18(*a*), the output angle

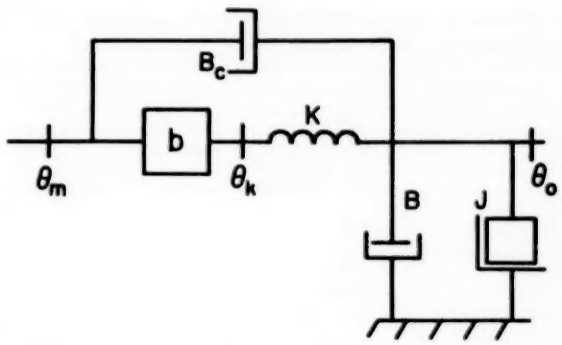


FIG. 21 ONE METHOD OF STABILIZING SYSTEMS WITH BACKLASH

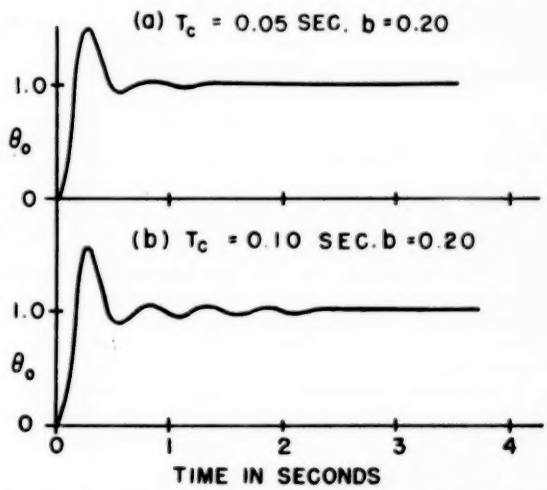


FIG. 22 RESPONSE TO UNIT-STEP INPUT; SYSTEM STABILIZED BY PARALLEL VISCOUS DAMPER

undergoes a small-amplitude high-frequency oscillation when the system is disturbed.

With backlash present, sustained oscillation occurs for that value of $G(\alpha)$ which causes the open-loop transfer-function locus to intersect the -1 point. Thus, if

$$\frac{\theta_o(j\omega)}{E(j\omega)} = -1$$

Equation [54] may be written

$$\frac{Q(j\omega)}{j\omega(j\omega T_m + 1)} = -\frac{G(\alpha) - \left(\frac{\omega}{\mu}\right)^2}{G(\alpha)} \quad \ldots\ldots [55]$$

If sustained oscillations are not to develop, then the linear function

$$G(j\omega) = \frac{Q(j\omega)}{j\omega(j\omega T_m + 1)} \quad \ldots\ldots [56]$$

must not intersect the locus of points described by

$$H(\alpha, j\omega) = -\frac{G(\alpha) - \left(\frac{\omega}{\mu}\right)^2}{G(\alpha)} \quad \ldots\ldots [57]$$

If a family of loci are plotted in the complex $G(j\omega)$-plane, each locus for constant $G(\alpha)$ and with the frequency ratio ω/μ the variable parameter, it will be found that all loci coincide. The loci of $H(\alpha, j\omega)$, defined as the "backlash stability line," begins at the point -1, for $\omega/\mu = 0$, and extends along the real axis to positive infinity. This backlash line is illustrated in Fig. 23(a) where the value of ω/μ at any point depends upon the particular value of $G(\alpha)$ in its possible range $0 \leq G(\alpha) \leq 1.0$.

It immediately follows from the foregoing that if the locus of the linear transfer function $G(j\omega)$ intersects the real axis, between the origin and the -1 point, sustained oscillations in the presence of backlash will occur.

A third-order transfer-function locus is shown in Fig. 23(b). The locus intersects the negative real axis at distance d from the point -1, at frequency ω_0. Sustained oscillations will then occur at the frequency ω_0. By Equation [55]

$$G(\alpha) = \frac{\left(\frac{\omega_0}{\mu}\right)^2}{d} \quad \ldots\ldots [58][4]$$

For this value of $G(\alpha)$, the ratio $b/2A_m$ is obtained from Fig. 8, and if the backlash b is known, the amplitudes of oscillation may be computed.

By Equation [44], letting $T \to \infty$ for this case, the amplitude of the output angle is

$$\theta_o(j\omega_0) = -\left(\frac{\mu}{\omega_0}\right)^2 G(\alpha) A_m \quad \ldots\ldots [59]$$

By Equation [45] the amplitude of the servomotor angle is

$$\theta_m(j\omega_0) = A_m + \theta_o(j\omega_0) \quad \ldots\ldots [60]$$

Eliminating $G(\alpha)$ from these three relations

[4] The undamped natural frequency of the load $\mu = \sqrt{K/J}$ must always be very great so that

$$\frac{(\omega_0/\mu)^2}{d} < 1.0$$

otherwise complete instability would result.

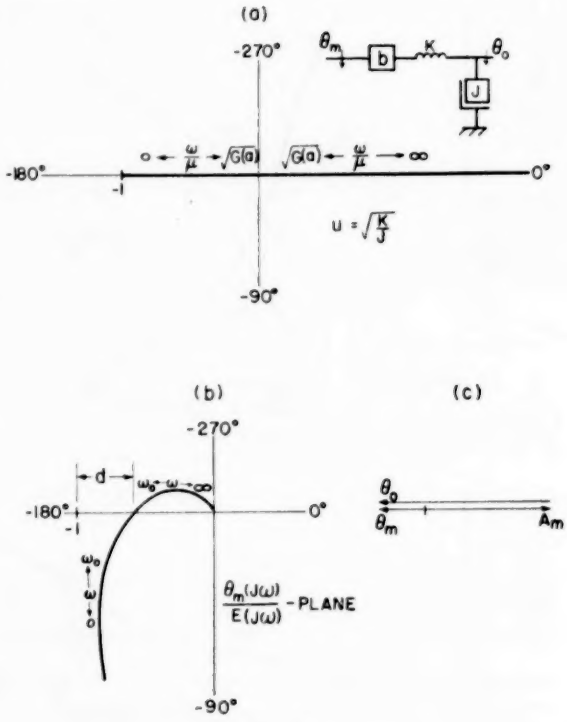


Fig. 23

(a, Backlash stability line. b, Third-order transfer function locus. c, Relative amplitude and phase of servomotor and output angles.)

$$\left.\begin{aligned} \theta_o(j\omega_0) &= -\frac{A_m}{d} \\ \theta_m(j\omega_0) &= -A_m\left(\frac{1}{d} - 1\right) \end{aligned}\right| \quad \ldots\ldots [61]$$

The in-phase oscillations just predicted were obtained on the differential analyzer as shown in Figs. 18(b), (c), and (d).

The ratio

$$\frac{b}{2\theta_o} = \frac{b}{2A_m}\frac{1}{G(\alpha)\left(\frac{\mu}{\omega_0}\right)^2}$$

computed for these cases is plotted versus system velocity constant in Fig. 24. The actual values obtained on the differential analyzer are indicated by points A, B, and C.

The foregoing analysis indicates that the frequency of sustained oscillation is equal to the frequency at which the linear function $G(j\omega)$ intersects the negative real axis. The frequency is therefore independent of system gain or velocity constant, since altering the gain merely linearly expands or contracts the open-loop transfer-function locus. Evidence supporting this conclusion is given in Fig. 18(c), where doubling the open-loop gain did not change the actual frequency of oscillation appreciably.

The dependency of the amplitude of sustained oscillation on the ratio of crossover frequency ω_0 to the load undamped natural frequency μ is shown in Fig. 25. Here the ratio of backlash b to peak-to-peak output oscillation amplitude $2\theta_o$ is plotted versus ω_0/μ for constant values of gain margin d. (Gain margin is frequently expressed as $20 \log_{10} d$ in decibels.) As depicted in the foregoing plots, for example, if $d = 0.40$, the amplitude of oscil-

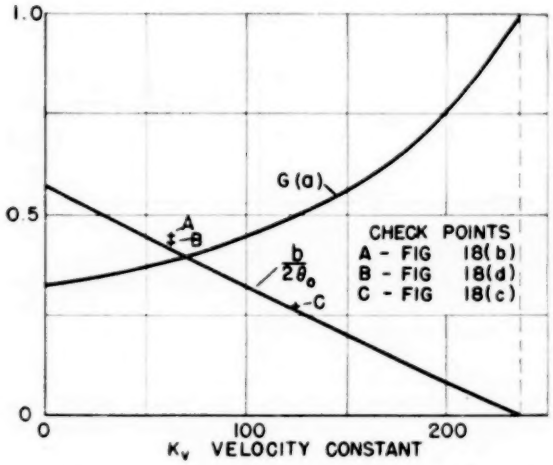


FIG. 24 DEPENDENCY OF AMPLITUDE OF OSCILLATION ON SYSTEM VELOCITY CONSTANT

lation θ_o becomes extremely large as $\omega_0/\mu \rightarrow 0.633$. As $\mu \rightarrow \infty$, $\omega_0/\mu \rightarrow 0$, the amplitude of oscillation decreases but never becomes zero. Maximum rigidity of mechanical components, nevertheless, is desirable. These characteristics, applicable to all systems having backlash of the type described, were computed from Equations [58] and [59] with the aid of Fig. 8.

CONCLUSIONS

The frequency-response method of analysis greatly facilitates the study of the effect of dead band in closed-loop systems. The results obtained here, although approximate, are in a form which enables the designer to determine means of suppressing or preventing the small-amplitude oscillations which occur.

The frequency-response method of analysis is well suited for the study of automatic control systems where some nonlinearity may produce sustained oscillations. The approximations generally made in such analyses are justified on the basis that automatic-control systems are essentially low-pass filters, resulting in an input signal to the nonlinear element nearly sinusoidal in form. The analysis presented here, however, shows that this is not necessarily a limitation, although this fact does simplify the analytical procedure. As long as the signals are of the periodic type, and the form of the input wave to the nonlinear element is known, the fundamental frequency-transfer ratio usually can be determined. The complexity of analysis is reduced by approximating the input wave by a simple Laplace transformable function.

In cases where the input wave to the nonlinear element, as in the case of dead band, does not change its form appreciably with frequency, the fundamental frequency-transfer ratio is truly amplitude-variant. If, however, the input wave changes form with frequency in such a way that phase shift of the fundamental frequency is produced, then the fundamental frequency-transfer ratio may become frequency-variant as well as amplitude-variant. The degree to which this occurs will depend upon the frequency-response characteristics of the entire closed-loop system. Fortunately, this effect is usually small.

ACKNOWLEDGMENTS

The author is indebted to Mr. Charles Concordia of the General Electric Company, Schenectady, N. Y., for many helpful suggestions.

BIBLIOGRAPHY

1 "The Analysis and Synthesis of Linear Servomechanism," by A. C. Hall, Technology Press, Cambridge, Mass., 1943.

2 "Principles of Servomechanisms," by G. S. Brown, and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948.

3 "Fundamental Theory of Servomechanisms," by L. A. MacCall, D. Van Nostrand Company, New York, N. Y., 1945.

4 "Servomechanisms and Regulating Systems Design," by H. Chestnut and R. W. Mayer, John Wiley & Sons, Inc., New York, N. Y., 1951.

5 "Stabilization of a Servomechanism Subject to Large Amplitude Oscillation," by E. S. Sherrard, Trans. AIEE, *Applications and Industry*, vol. 71, part 2, 1952, pp. 312–324.

6 "Approximate Frequency-Response Methods for Representing Saturation and Dead Band," by H. Chestnut, published in this issue, pp. 1345–1363.

7 "A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, 1950, pp. 270–284.

8 "The Effects of Backlash and of Speed-Dependent Friction on the Stability of Closed-Cycle Control Systems," by A. Tustin, JIEE, part IIA, 1947, pp. 143–151.

9 "Sinusoidal Analysis of Feedback-Control Systems Containing Non-Linear Elements," by E. C. Johnson, Trans. AIEE, *Applications and Industry*, vol. 71, part 2, 1952, pp. 169–181.

10 "Effect of Prime-Mover Speed-Governor Characteristics on Power System Frequency Variations and Tie-Line Power Swings," by C. Concordia, S. B. Crary, and E. E. Parker, Trans. AIEE, vol. 60, 1941, pp. 559–567.

11 "The Dynamics of Automatic Controls," by R. C. Oldenbourg and H. Sartorius; translated by H. L. Mason, THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS, New York, N. Y., 1948.

12 "Transients in Linear Systems," by M. F. Gardner and J. L. Barnes, John Wiley & Sons, Inc., New York, N. Y., 1942.

Appendix

LIMIT OF STABILITY—BACKLASH AND INERTIA

As the undamped natural frequency μ of the load becomes extremely large, oscillations resulting from backlash, although reduced in amplitude, still occur. This is illustrated in Fig. 25. In this case, $\omega_0/\mu \rightarrow 0$, and $G(\alpha) \rightarrow 0$, both simultaneously. Zero is merely a limit; both ω_0/μ and $G(\alpha)$ for practical reasons always will be slightly greater than zero. The backlash-line shown in Fig. 23(a) is still applicable, and the amplitudes of fundamental-frequency oscillation are given by Equations [61]. In this case, however, as $G(\alpha)$ becomes very small $2A_m/b \rightarrow 1.0$ (see Fig. 8). Equations [61] then simplify to

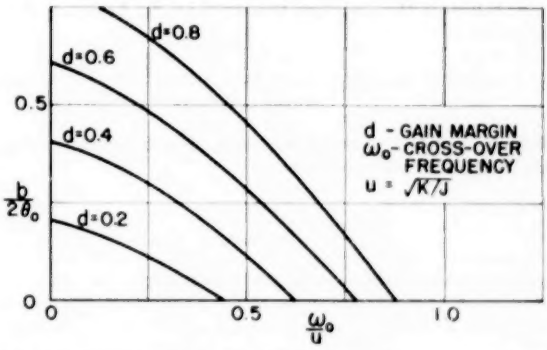


FIG. 25 DEPENDENCY OF AMPLITUDE OF OSCILLATION ON SYSTEM GAIN MARGIN AND RATIO OF CROSSOVER FREQUENCY TO NATURAL FREQUENCY OF LOAD

$$\left.\begin{aligned}\theta_o(j\omega_0) &= -\frac{b}{2d}\\ \theta_m(j\omega_0) &= -\frac{b}{2}\left(\frac{1}{d}-1\right)\end{aligned}\right| \quad \ldots\ldots [62]$$

It may be concluded in this case that as the mechanical parts become extremely rigid, the amplitude of oscillation depends solely on the backlash b and the gain margin d. The frequency of oscillation ω_0 is that frequency for which the linear transfer-function locus $G(j\omega)$ intersects the negative real axis between the origin and the —1 point.

With sustained oscillations occurring, the servomotor angle θ_m and output angle θ_o are in time phase, since $d < 1.0$. Their velocities therefore are in time phase. During the course of oscillation, when the two velocities are not zero, the backlash zone is being traversed. With no damping present, the velocity of the output member is constant during the traversing of backlash. Backlash is taken up simultaneously as the velocity of the input member goes to zero. Reverse motion is imparted to the output member at the same velocity existing before impact (assuming an elastic collision and high servomotor inertia referred to the load side of the dead band). The servomotor velocity then may be approximately sinusoidal, and the velocity of the output member is a square wave.

If this idealized type of system is of second order or less, complete instability occurs with or without backlash. This may be explained by the fact that in such cases the transfer-function locus does not intersect the negative real axis, and a cascaded function of the form

$$\frac{\theta_o}{\theta_m}(j\omega) = \frac{1}{1-\left(\frac{\omega}{\mu}\right)^2} \quad \ldots\ldots [63]$$

will produce enclosure of the —1 point in the complex plane.

Limit of Stability—Backlash, Spring, and Viscous Damping

In cases where the effect of load inertia is negligible, such that $T = J/B \to 0$, the open-loop transfer function, Equation [53], reduces to

$$\frac{\theta_o(j\omega)}{E(j\omega)} = Q(j\omega)\,\frac{1}{j\omega(j\omega T_m + 1)}\,\frac{G(\alpha)}{j\omega T_k + G(\alpha)} \quad \ldots [64]$$

where $T_k = B/K$. If sustained oscillations are not to develop, the linear transfer-function locus

$$G(j\omega) = \frac{Q(j\omega)}{j\omega(j\omega T_m + 1)} \quad \ldots\ldots [65]$$

must not intersect the locus of points defined by

$$H_f(\alpha, j\omega) = -1 - j\,\frac{\omega T_k}{G(\alpha)} \quad \ldots\ldots [66]$$

Another backlash-stability line is thus defined.

The locus of $H_f(\alpha, j\omega)$ is a straight line in the complex $G(j\omega)$-plane as shown in Fig. 26(a). Transfer-function loci of systems designed for high performance usually will intersect this backlash line. A stable system will have two points of intersection. The open-loop transfer-function locus of a position-control system of the third order is sketched in Fig. 26(b). The frequencies at the intersection points are ω_1 and ω_2 where $\omega_1 > \omega_2$. Two possible values of $G(\alpha)$ are indicated. Thus

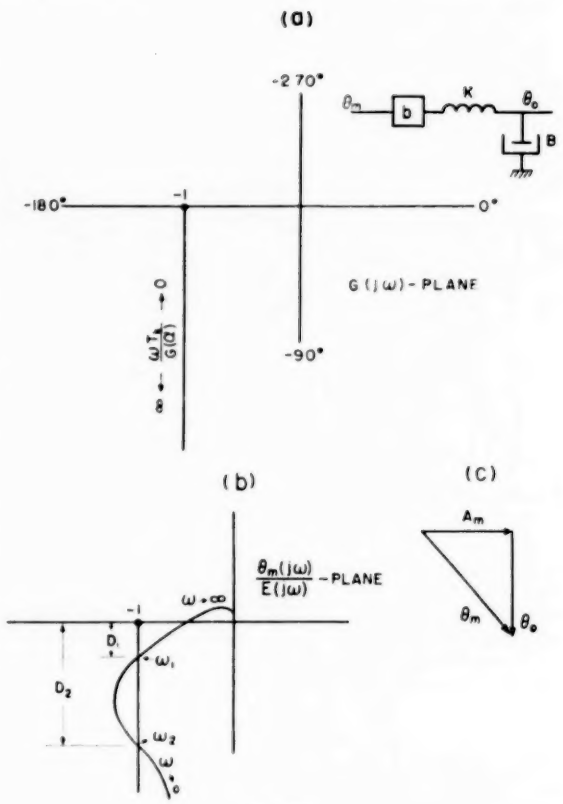


Fig. 26

(a. Backlash stability line for load characterized by spring action and viscous damping. b. Third-order transfer function locus. c. Relative amplitude and phase of servomotor and load positions.)

$$G_1(\alpha) = \frac{\omega_1 T_k}{D_1} \quad \ldots\ldots [67]$$

$$G_2(\alpha) = \frac{\omega T_k}{D_2} \quad \ldots\ldots [68]$$

Now, since $\omega_1 > \omega_2$ and $D_1 < D_2$, it follows that $G_1(\alpha) > G_2(\alpha)$. The larger value of $G(\alpha)$ corresponds to larger amplitudes of oscillation, Fig. 8. Therefore, for amplitudes of oscillation in the range $D_2 > D > D_1$, the system is unstable and tends to oscillate with increasing amplitude. The stable point of oscillation then is at the higher frequency ω_1 corresponding to larger amplitudes of oscillation. The fundamental frequency amplitudes of oscillation, Equations [44] and [45], reduce to

$$\theta_o(j\omega) = -j\,\frac{G_1(\alpha)}{\omega_1 T_k}\,A_m \quad \ldots\ldots [69]$$

$$\theta_m(j\omega) = A_m + \theta_o(j\omega) \quad \ldots\ldots [70]$$

On eliminating $G_1(\alpha)$ between Equations [67] and [69]

$$\theta_o(j\omega) = -j\,\frac{A_m}{D_1} \quad \ldots\ldots [71]$$

In cases where the deflection of the equivalent spring is small compared to the dead band

$$G_1(\alpha) \to 0 \text{ and } A_m \to \frac{b}{2}$$

Equations [70] and [71] then become

$$\theta_o(j\omega) = -j\frac{b}{2D_1} \quad [72]$$

and

$$\theta_m(j\omega) = \frac{b}{2}\left[1 - j\frac{1}{D_1}\right] \quad [73]$$

The fundamental frequency amplitude of oscillation is thus directly proportional to the dead band b, and inversely proportional to the distance between the -1 point and the nearest point of intersection of the backlash stability line.

The foregoing, of course, assumes that the phase shift across the dead-band element is zero for fundamental frequencies. As the equivalent spring becomes infinitely stiff, the approximation in the analysis becomes greater. The correct position of the backlash line is very well predicted nevertheless. A comparison with the backlash line computed by Johnson,[8] the characteristic illustrated in Fig. 1(*b*), indicates that the result obtained here is somewhat pessimistic.

Stability Limit—Dead Zone

If the insensitive zone may be described by the characteristic shown in Fig. 1(*a*), the open-loop transfer function becomes

$$\frac{\theta_o(j\omega)}{E(j\omega)} = \frac{Q(j\omega)}{j\omega(j\omega T_m + 1)}G(\alpha) \quad [74]$$

Proceeding as before, it follows that if sustained oscillations are not to develop, the frequency-variant transfer-function locus must not intersect the locus of points defined by

$$H_2 = -\frac{1}{G(\alpha)} \quad [75]$$

Since $0 < G(\alpha) < 1.0$, the dead-zone stability line lies on the negative real axis between -1 and $-\infty$, Fig. 27(*a*). Only systems of the conditionally stable type will intersect this line. The transfer-function locus of a conditionally stable system is sketched in Fig. 27(*b*). Again, the intersection point at the higher frequency corresponds to larger amplitudes of oscillation, Fig. 8. Points between the frequencies ω_1 and ω_2 corresponding to lower amplitudes of oscillation are enclosed by the locus. The system is unstable in this region; hence any disturbance will cause sustained oscillation at the frequency ω_1. $G(\alpha)$ is determined by the distance between the -1 point and the nearest point of intersection of the dead-zone stability line, hence

$$G_1(\alpha) = \frac{1}{D_1 + 1} \quad [76]$$

In this case $A_m = \theta_m$; therefore Fig. 8 represents the relationship between $1/(D_1 + 1)$ and $b/2\theta_m$, where θ_m is the fundamental-frequency amplitude of oscillation.

Dead-Band Transfer Ratio for a Sinusoidal Input Wave

The input wave, Fig. 7, is defined in the interval $0 \leq t \leq T/2$ by

$$\alpha_0(t) = A_m \sin \omega t\left[u(t) - u\left(t - \frac{T}{2}\right)\right] \quad [77]$$

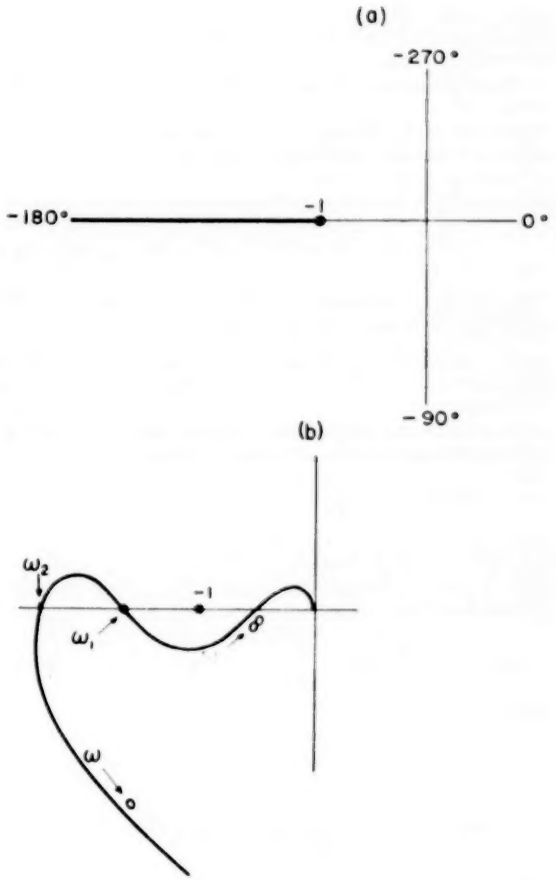


FIG. 27

(*a*. Dead zone stability line. *b*. Transfer function locus of conditionally stable system. *c*. Relative amplitude and phase of dead-zone input and output variables.)

where $u(t)$ is the unit step function, and $u(t - T/2)$ is the unit step function translated to the right by $T/2$. Similarly, for $0 \leq t \leq T/2$, the output of the dead-band element is defined by

$$\beta_0(t) = A_m \sin \omega t\, u(t - t_1) - A_m \sin \omega t\, u(t - t_2) - \frac{b}{2}[u(t - t_1) - u(t - t_2)] \quad [78]$$

The input wave is sectioned at t_1 and t_2, and the difference between the two resulting waves is taken. On subtracting $b/2$ in the designated interval, the output function is obtained.

Making use of the theorem of real translation from Laplace transform theory (12) and the transform pairs

$$\mathcal{L} \sin \omega t = \frac{\omega}{s^2 + \omega^2}, \quad \mathcal{L} \cos \omega t = \frac{s}{s^2 + \omega^2}$$

the transforms of the foregoing two half-period waves are

$$A_0(s) = \frac{A_m \omega}{s^2 + \omega^2}(1 + e^{-T/2s}) \quad [79]$$

and

[8] Reference (9), fig. 6.

$$B_0(s) = \frac{A_m}{s^2 + \omega^2}\left[(s \sin \omega t_1 + \omega \cos \omega t_1)e^{-t_1 s} - (s \sin \omega t_2 + \omega \cos \omega t_2)\, e^{-t_2 s}\right] - \frac{b}{2s}\,(e^{-t_1 s} - e^{-t_2 s}) \quad [80]$$

Since $b/2 = A_m \sin \omega t_1$, $\sin \omega t_1 = \sin \omega t_2$, and $\cos \omega t_1 = -\cos \omega t_2$, Equation [80] reduces to

$$B_0(s) = \frac{A_m \omega}{s^2 + \omega^2}\left[(e^{-t_1 s} + e^{-t_2 s}) \cos \omega t_1 - \frac{\omega}{s}\,(e^{-t_1 s} - e^{-t_2 s}) \sin \omega t_1\right] \quad [81]$$

The half-period transform ratio $B_0(s)/A_0(s)$ is identical to the ratio of the transforms of the two wave trains, therefore

$$\frac{B_0(s)}{A_0(s)} = \frac{(s \cos \omega t_1 - \omega \sin \omega t_1)e^{-t_1 s} + (s \cos \omega t_1 + \omega \sin \omega t_1)e^{-t_2 s}}{s(1 + e^{-T/2s})} \quad [82]$$

As $s \to j\omega$, Equation [82] reduces to the indeterminate form 0/0. Applying L'Hospital's rule and evaluating yields

$$\frac{B_0(j\omega)}{A_0(j\omega)} = \frac{1}{\pi}\,[\omega(t_2 - t_1) - 2 \sin \omega t_1 \cos \omega t_1] \quad [83]$$

This ratio is not a function of frequency, since by Fig. 7, $\omega(t_2 - t_1) = \pi - 2\omega t_1$ and $\sin \omega t_1 = b/2A_m$. Thus Equation [83] may be written

$$G(\alpha) = 1 - \frac{2}{\pi}\left[\sin^{-1} \frac{b}{2A_m} + \frac{b}{2A_m}\sqrt{1 - \left(\frac{b}{2A_m}\right)^2}\right] \quad [84]$$

where

$$G(\alpha) = \frac{B_0(j\omega)}{A_0(j\omega)} \quad [85]$$

Dead-Band Transfer Ratio for a Triangular Wave Input

The first half cycle of the input wave, Fig. 9, is defined by

$$\alpha_0(t) = mt\, u(t) - 2m\left(t - \frac{T}{4}\right) u\left(t - \frac{T}{4}\right) + m\left(t - \frac{T}{2}\right) u\left(t - \frac{T}{2}\right) \quad [86]$$

where m is the time rate of change of $\alpha(t)$. Taking the Laplace transform

$$A_0(s) = \frac{m}{s^2}\,(1 - 2e^{-T/4s} + e^{-T/2s}) = \frac{m}{s^2}\,(1 - e^{-T/4s})^2 \quad [87]$$

Similarly, the first half cycle of the output wave is

$$\beta_0(t) = m(t - t_1)\, u(t - t_1) - 2m\left(t - \frac{T}{4}\right) u\left(t - \frac{T}{4}\right) + m(t - t_2)\, u(t - t_2) \quad [88]$$

Laplace transforming

$$B_0(s) = \frac{m}{s^2}\,(e^{-t_1 s} - 2e^{-T/4s} + e^{-t_2 s}) \quad [89]$$

The ratio of Laplace transforms

$$\frac{B_0(s)}{A_0(s)}$$

is identical to the ratio of Laplace transforms of the two wave trains. Thus

$$\frac{B_0(s)}{A_0(s)} = \frac{e^{-t_1 s} - 2e^{-T/4s} + e^{-t_2 s}}{(1 - e^{-T/2s})^2} \quad [90]$$

as $s \to jn\omega$

$$\frac{B_0(jn\omega)}{A_0(jn\omega)} = 1 - \frac{\sin n\omega t_1}{\sin n\,\frac{\pi}{2}} \quad [91]$$

In the foregoing step, use is made of the identities $\cos n\omega t_1 = -\cos n\omega t_2$, $\sin n\omega t_1 = \sin n\omega t_2$, $\omega T = \pi$, where n = odd integer. A further simplification is made by noting in Fig. 9 that $m(T/4) = A_m$ and $mt_1 = b/2$. Therefore

$$n\omega t_1 = \frac{n\pi}{2}\,\frac{b}{2A_m}$$

with this change

$$\frac{B_0(jn\omega)}{A_0(jn\omega)} = G(\alpha, jm\omega) = 1 - \frac{\sin\left(\frac{n\pi}{2}\,\frac{b}{2A_m}\right)}{\sin n\,\frac{\pi}{2}} \quad [92]$$

The transfer ratio for the fundamental frequency, $n = 1$, is

$$G(\alpha, j\omega) = 1 - \sin \frac{b}{2A_m}\,\frac{\pi}{2} \quad [93]$$

Discussion

A. Tustin.[6] This paper and one by Harold Chestnut[7] add significantly to previous studies that have been made of the approximate effect of the typical nonlinearities usually present in servosystems, in terms of responses to sinusoidal variations.

What is now required is to draw from these studies the appropriate conclusions as to the features required in a design in order to meet a specification for accuracy in the most convenient way. This is not an obvious or simple matter, because several different kinds of nonlinearity are usually present together and because the best arrangements depend also on the amount and nature of the disturbances present.

It is proposed to take for illustration the case of position controls, of the larger powers, and in cases where high accuracy is required. Characteristically, in such systems, the following major nonlinearities occur:

1 Torque and speed limitation are present, not as defects, but as requirements. These usually are provided a combination of three features; namely, (*a*) limiting by saturation the output or signal at the output stage of the thermionic amplifier, (*b*) limiting the voltage of the final machine stage, and (*c*) providing a feedback from output current to a point between these two saturating elements. This combination limits both torque and speed.

2 The friction stiction characteristic of the load imposes requirements to avoid jerky operation at low speeds.

3 Backlash, with resilience, is present.

[6] Professor, University of Birmingham, Birmingham, England.

[7] "Approximate Frequency-Response Methods for Representing Saturation and Dead Band," by Harold Chestnut, published in this issue, pp. 1345–1363.

The principles for good design that would seem to result from this situation are the following:

(*a*) The system should be so designed that normal operation, for which the specified accuracy is to be obtained, should not involve the speed or torque limitations, because the error increases extremely rapidly if these limits are exceeded. The system then operates as if the speed and torque limits did not exist, so far as concerns the normal regular operation as a servo. Incidentally, however, the presence of the feedback from output current makes the system operate, from most practical points of view, as a torque control or second-order servo. It now requires phase advance if it is to be stable.

(*b*) The real incidence of torque and speed limitations is when running into alignment from large initial misalignments, i.e., the well-known phenomena of long "settling time" and repeated overshoots. For this defect to be avoided a very large first-derivative term must appear in the signal effective at low frequencies, and prior to the saturating tube stage, so as to reverse the torque a sufficient time before alignment is reached. Since large-amplitude oscillation sometimes may occur, and the torque limitation then has much the same effect as a decrease in loop gain, it is desirable to avoid designs having "conditional stability." This also is assured if there is a large phase lead at low frequencies, and this point of view is really saying the same thing in another way.

(*c*) The presence of a saturable thermionic-tube stage is of critical significance, not because it affects the accuracy on normal operation, since such saturation can be avoided in normal operation by the basic design, but because such systems are usually subject to high-frequency disturbances mixed with the input. These high-frequency components are not to be followed, and would often be unimportant were it not for the fact that, amplified by the presaturation lead networks, they saturate and block the saturable stage, reducing the effective gain for legitimate signals.

A logical escape from these conflicting difficulties would be to limit the extra gain of the modifying networks preceding the saturable stage, or, ultimately, to include there a filter or high-frequency attenuator. If this is done the only place left for the provision of the necessary further shaping of the frequency locus is the postsaturation stages, i.e., usually the rotary amplifier or its equivalent.

It is this combination of circumstances that makes it desirable to develop rotary-amplifier combinations, that contribute phase advance in the region of critical system frequencies (say 2 to 8 cps).

It has sometimes been asserted that efforts to improve the frequency response characteristics of rotary amplifiers are unnecessary, because their lags can be compensated by modifying networks in the thermionic stages. This would be true if the systems were linear, but it is quite untrue if torque limitation and input disturbance are both present.

It also may be shown that both backlash and the requirement for smooth operation at very low speeds also impose a requirement for phase advance, and if this cannot be provided prior to the saturable stage because of the need to filter out noise, it must be provided later in the sequence.

An outstanding need, therefore, and one which it is not at all impossible to meet, is the provision of greatly improved rotary amplifiers, not only having a much lower ratio of time-constant to power amplification than those at present available, but, by the provision of suitable local feedbacks, contributing significantly to the phase advance that is essential in these effectively second-order systems. Work with this aim is being done both at Birmingham and at the Massachusetts Institute of Technology.

R. L. Cosgriff.[8] The author is to be congratulated for his interesting development of the well-known dead-band or backlash problem. This problem has been widely recognized and studied. The method proposed by Dr. Johnson (reference 9 of the paper) was an excellent beginning. However, because of the limitations imposed in this method, the analytical results obtained with this method often do not agree too well with measured results. This deviation between computed and measured results is primarily due to the inertia of the servo load. The method proposed by the author is certainly an important contribution for it does not have the limitations of this earlier frequency-response method.

Dead band and coulomb friction are two types of nonlinearities that cannot be handled using small signal theory. In fact, systems with this type of nonlinearity become linear as the amplitudes of the signals approach infinity. As this is the case they adversely affect the performance of servo systems when the input signals are small or zero. As the author indicates, it is difficult to eliminate oscillations caused by this type of nonlinearity by linear means. It has been the writer's opinion that the most effective method for eliminating oscillation caused by backlash is by means of coulomb friction. The writer would like to ask if the author has considered this method for stabilizing the systems he has studied.

The writer also would like to suggest that the author extend the scope of his paper by deriving equations expressing the value of $G(\alpha)$ and the frequency of oscillation ω as functions of the system parameters. For example, if $Q(j\omega)$ is a constant Q_0, it can be shown that a condition for steady-state oscillation is

$$G(\alpha) = \frac{\dfrac{K}{\mu^2 T} - Q_0}{K\left(\dfrac{T_m K}{\mu^2} - T_m - \dfrac{k}{\mu^2}\right)}$$

where

$$K = \frac{\mu^2 T + k}{T + T_m}$$

and the frequency of oscillation is given by

$$\omega = \sqrt{KG(\alpha)}$$

It seems likely that similar expressions can be derived for $G(\alpha)$ when $Q(j\omega)$ is a function of frequency.

Author's Closure

Professor Tustin outlines quite clearly the practical problems which are of primary consideration in the design of a position control of high accuracy. In outlining the features of a good design he chooses for illustration a position control where the power requirement is large. In so doing he effectively points out the fundamental limitations of position control.

As the power requirement increases, special attention must be given to maintaining effectively short time constants throughout the system. Any closed-loop system having an element of pure integration in the main loop, as is characteristic of position controls, cannot tolerate a single "long" time constant if the velocity constant is to be set high, say, 10 to 50 sec^{-1}. Contrarily, in regulating systems where pure integration is absent, a single long time constant not only is desirable but is required.

In d-c motor generator sets the long time constant usually occurs in the generator field circuit. To offset the effect of this time

[8] Assistant Professor, Department of Electrical Engineering, Ohio State University, Columbus, Ohio.

constant, the rotary amplifier supplying the field must be capable of developing several times normal operating voltage and, in addition, must be extremely fast. In the presence of a major disturbance a phase-leading network preceding the rotary amplifier facilitates the driving of the amplifier toward its ceiling and back to normal voltage again in an appropriate interval of time. If compensation by such means is inadequate, it may be necessary to shorten the generator field time constant by employing fewer turns on the generator field. This change then would necessitate the using of a rotary amplifier of higher power rating. High accuracy and acceptable relative stability can be achieved more economically, if a maximum of performance is imposed on the rotary amplifier.

Thus the need of phase-advancing networks and components in position-control systems of high power is normally greater than that encountered in low power systems. System stability is not the only consideration, of course. Fast response is required if the torque or motor armature current is to be limited effectively following a sudden change in load. The development work described by Professor Tustin, in which a phase-advancing characteristic is incorporated into the design of the rotary amplifier, is interesting indeed.

Additional phase advance may be required in the presence of dead band or backlash, if sustained small-amplitude oscillations are to be minimized or eliminated. For example, Equations [61] and [62] of the paper indicate that it is desirable for the linear transfer-function locus (Equation [56] or Equation [65]) not to penetrate appreciably the region enclosed by a circle of unit radius having the -1 point as center.

In answer to the question asked by Professor Cosgriff, the author has not considered the effect of coulomb friction in combination with dead band. It is known, however, that coulomb friction also can cause oscillation, or an insensitive zone of response.[9] If sustained oscillation was eliminated by such means, it would appear to be undesirable because of the accompanying poor response to small signals.

The author wishes to thank both Professor Tustin and Professor Cosgriff for their contribution.

[9] "Coulomb Friction in Feedback Control Systems," by V. B. Haas, Jr., Trans. AIEE, vol. 72, part II, 1953, pp. 119–123.

Experimental Flight Methods for Evaluating Frequency-Response Characteristics of Aircraft

By G. A. SMITH[1] AND W. C. TRIPLETT,[1] MOFFETT FIELD, CALIF.

This paper deals mainly with methods and techniques for obtaining airplane frequency-response characteristics from transient flight measurements. After a review of the basic airplane equations of motion, the theoretical transfer function which defines the "short-period" pitch response to an elevator deflection is derived. A numerical example is used to illustrate the evaluation of an analytical transfer function from a graphical frequency response by a method which involves the use of graphical aids in conjunction with an analog computer. Finally, the effects of variations in flight speed and altitude on the airplane response are considered.

Nomenclature

The following nomenclature is used in the paper:

I_x, I_y, I_z	=	moments of inertia about X, Y, and Z-axes, slug-ft^2
I_{xz}	=	product of inertia, slug-ft^2
K	=	gain constant
L, M, N	=	rolling, pitching, and yawing moments (Fig. 2), ft-lb
S	=	wing area, sq ft
V	=	velocity, fps
X, Y, Z	=	forces coincident with axes shown in Fig. 2, lb
c	=	mean aerodynamic chord, ft
m	=	mass, slugs
q	=	pitching velocity, radians per sec (rps)
s	=	Laplace operator
t	=	time, sec
i	=	$\sqrt{-1}$
α	=	angle of attack, radians
δ	=	elevator deflection, radians (except as noted)
θ	=	pitch angle, radians
ρ	=	air density, slugs/cu ft
ζ	=	damping ratio
φ	=	phase angle, deg
ω	=	angular frequency, rps
ω_n	=	undamped natural frequency, rps
τ	=	time constant, sec

Dot over quantity indicates derivative with respect to time.

Introduction

The extremely wide range of speeds and altitudes at which the latest aircraft operate leads to serious problems in stability and control. The difficulty experienced by the human pilot in coping with these problems has contributed greatly to the emphasis now being placed on the automatic interceptor and the guided missile which depend on high-performance autopilots for precise maneuvering control during combat operations. For the analysis of these complex systems which involve the dynamics of the airframe, the frequency-response methods of servomechanism theory have been applied successfully.

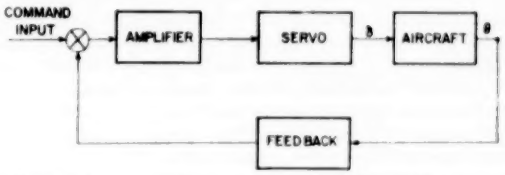


Fig. 1 Simplified Airplane-Autopilot Block Diagram

A simplified block diagram of a typical airplane-autopilot control loop is shown in Fig. 1. The analysis of such a closed-loop system requires a knowledge of the dynamic behavior of each element of the loop or, more precisely, the transfer functions which define the ratio of output to input for each element. The transfer functions of the airplane define, for example, the manner in which the airplane responds to motions of the control surfaces. Unfortunately, the aircraft transfer functions vary markedly as the operating altitude and speed of the aircraft change, and considerable effort must be expended in determining the nature of these variations. This feature distinguishes the airplane-autopilot problem from the ordinary servomechanism problem, and means that gain changers and variable stabilizing networks must be included in the autopilot to insure satisfactory performance of the complete loop as flight conditions change. The problem is complicated further by the exacting performance required of maneuvering-type autopilots which must control a fighter aircraft during combat maneuvers. This is in contrast to the situation that existed before World War II, when autopilots were used primarily for pilot relief on aircraft that operated over more limited ranges of speed and altitude.

A similar analysis problem exists in some piloted aircraft where it has been necessary to modify the dynamic responses of the airplane by means of servo devices such as yaw and pitch dampers so that the pilot may control the airplane properly under all flight conditions. This problem arises primarily from the inherently low damping and high natural frequency which accompany high-speed high-altitude flight. The problem of making the combination of aircraft and damper satisfactory for piloted operation is, in effect, a closed-loop problem similar to that for the airplane-autopilot combination described previously, except that the human pilot is now an element in the loop. Because there is no entirely satisfactory transfer function to describe the human response, the problem involves proper tailoring of the dynamic characteristics of the damper so that when combined with the airplane, the resulting airplane-damper combination will have a transfer function that is known by previous experience to be acceptable to the human pilot.

The foregoing has indicated some of the needs for airplane fre-

[1] Aeronautical Research Scientist, Ames Aeronautical Laboratory, NACA.

Contributed by the Instruments and Regulators Division and presented at the Semi-Annual Meeting, Pittsburgh, Pa., June 20–24, 1954, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters on February 15, 1954. Paper No. 54—SA-3.

quency-response data in coping with stability and control problems of high-performance aircraft. Papers by Bollay (1)[2] and Milliken (2) give a more complete background and summarize some of the progress made in this general field.

The present paper is concerned mainly with methods for evaluating the dynamic-response characteristics of the airplane. This information in transfer-function form can be calculated by substituting known values of the physical parameters into the differential equations that describe the airplane motions. In addition to the usual mass and inertia constants, these parameters consist of terms which relate the aerodynamic forces and moments to the motions of the airplane. These aerodynamic terms, when expressed in dimensionless coefficient form, are called stability derivatives and may be estimated from theoretical studies and the results of wind-tunnel tests. In some cases this may be done with enough accuracy to give reliable predictions of the airplane transfer functions. However, it is generally advisable to use experimental flight tests in order to obtain sufficiently complete and accurate data for purposes of analysis. This is particularly true for flight in the transonic speed range where large and erratic variations of these parameters with Mach number cannot be predicted reliably from theory or from wind-tunnel tests.

Furthermore, there are cases where the customary "rigid-body" equations generally used do not define completely the aircraft motions; aeroelastic effects are a case in point. These are changes in flight-path stability which result from bending and twisting of an airplane structure while under the influence of inertia and aerodynamic loads. The occurrence of additional peaks in the experimental frequency-response amplitude curve at frequencies higher than the predicted rigid-body modes of the aircraft usually indicates the presence and degree of importance of structural modes in the system. The motions resulting from aeroelastic deformations may be of importance in the selection of a suitable autopilot and in the location of its sensing elements (3).

The effects of Mach-number variations and aeroelastic deformations on the airplane response are two of the practical difficulties that beset the purely theoretical approach and make it necessary to resort to flight methods to obtain adequate aircraft transfer functions. Flight tests also serve an additional purpose by enabling comparisons to be made with predicted results, thus aiding in the development of more refined prediction methods.

After a brief discussion of the aerodynamic theory underlying the response equations of the aircraft, the body of this paper will be devoted to a consideration of the determination of aircraft frequency response from the analysis of flight data.

Development of Theoretical Transfer Functions

An airplane in flight can be considered as a dynamic system possessing six degrees of freedom. It can translate along and rotate about each of the axes, X, Y, and Z shown in Fig. 2. Thus there are six equations which define the motions of the airplane and make possible the calculation of responses to various types of disturbing forces. The following expressions are obtained by equating inertia forces and moments (on the left-hand side) to aerodynamic forces and moments (on the right-hand side)

$$m(\dot{u} + qw - rv) = X \quad [1]$$

$$m(\dot{v} + ru - pw) = Y \quad [2]$$

$$m(\dot{w} + pv - qu) = Z \quad [3]$$

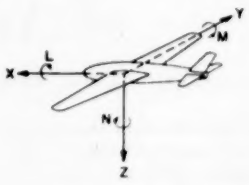
$$I_x\dot{p} + (I_z - I_y)qr - I_{xz}(\dot{r} + pq) = L \quad [4]$$

$$I_y\dot{q} + (I_x - I_z)rp - I_{xz}(r^2 - p^2) = M \quad [5]$$

$$I_z\dot{r} + (I_y - I_x)pq - I_{xz}(\dot{p} - qr) = N \quad [6]$$

The complete development of these equations may be found in any of several references (4, 5) and they are only briefly examined here.

It can be seen that the system represented by the equations is nonlinear and extremely complicated. Fortunately, however, the motions of the airplane are often restricted in such a manner that the system, for most practical purposes, becomes two independent linear systems, each with three degrees of freedom. Assume first that a disturbing force is applied so that the airplane is disturbed only in the plane of symmetry (the X-Z-plane). This disturbance may be an elevator-control deflection, a change in engine power, or an atmospheric gust. The resultant motions generally are referred to as the "longitudinal" or "pitch" responses of the airplane and consist of accelerations along the X and Z-axes and rotations about the Y-axis. Because of symmetry, there are no motions in any other plane.



AXIS	X	Y	Z
FORCE	X	Y	Z
MOMENT	L	M	N
LINEAR VELOCITY	u	v	w
ANGULAR VELOCITY	p	q	r

Fig. 2 Axis Notation and Definition of Symbols

If, on the other hand, the airplane is disturbed by a motion of the aileron or rudder, it will rotate about the X and Z-axes (roll and yaw) and translate along the Y-axis (sideslip). Motions of this type generally are referred to as "lateral" responses of the airplane. The complete equations indicate that lateral motions will induce secondary pitching motions as a result of inertia coupling. However, when the system is restricted to small displacements from level flight, the product terms that cause this pitch coupling are comparatively small and often can be neglected. The simplified equations for the two systems then become, for the longitudinal system

$$m(\dot{u} + qw) = X \quad [7]$$

$$m(\dot{w} - qu) = Z \quad [8]$$

$$I_y\dot{q} = M \quad [9]$$

and for the lateral system

$$m(\dot{v} + ru) = Y \quad [10]$$

$$I_x\dot{p} - I_{xz}\dot{r} = L \quad [11]$$

$$I_z\dot{r} - I_{xz}\dot{p} = N \quad [12]$$

These equations then define the complete motions of the airplane in the absence of cross coupling between longitudinal and lateral motions.

Turning our attention now to the longitudinal responses which result from a deflection of the elevator, we can make a further simplification. With the limitation of small disturbances, the change in flight velocity which accompanies the control deflection is generally very small, and the resultant X-force on the airplane is essentially zero, provided of course that there is no change in engine thrust. Thus the longitudinal system effectively reduces to two degrees of freedom in which the variable u can be replaced by the constant velocity V. For convenience, the normal velocity w can be expressed as $V\alpha$ where α is the angle of attack of the airplane, i.e., the angle in the plane of symmetry between the relative wind vector and the reference axis of the airplane. The longitudinal equations then become

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

$$mV(\dot{\alpha} - q) = Z \quad [13]$$

$$I_y\dot{q} = M \quad [14]$$

In order to solve for the responses q, α, or the normal acceleration $V(\dot{\alpha} - q)$, it is necessary to express the aerodynamic forces and moments, M and Z, in terms of α, q, and the elevator deflection δ. For small displacements, Z and M are primarily linear functions of α, $\dot{\alpha}$, q, and δ and Equations [13] and [14] may be written as

$$mV(\dot{\alpha} - q) = \frac{\partial Z}{\partial \alpha}\alpha + \frac{\partial Z}{\partial \dot{\alpha}}\dot{\alpha} + \frac{\partial Z}{\partial q}q + \frac{\partial Z}{\partial \delta}\delta \quad [15]$$

$$I_y\dot{q} = \frac{\partial M}{\partial \alpha}\alpha + \frac{\partial M}{\partial \dot{\alpha}}\dot{\alpha} + \frac{\partial M}{\partial q}q + \frac{\partial M}{\partial \delta}\delta \quad [16]$$

In accordance with our limitation of small displacements, the partial derivatives are constant for a particular velocity and flight altitude. In these equations α and δ are incremental angles and are the difference between the instantaneous values and the values for trimmed level flight.

We now have two linear differential equations with constant coefficients which can be solved simultaneously for α and q in terms of δ. For ease in writing, the force equation is divided by mV and the moment equation by I_y and a shorthand notation adopted so that the equations read

$$\dot{\alpha} - q = Z_\alpha\alpha + Z_\delta\delta \quad [17]$$

$$\dot{q} = M_\alpha\alpha + M_{\dot{\alpha}}\dot{\alpha} + M_q q + M_\delta\delta \quad [18]$$

Eliminating α from Equations [17] and [18] gives

$$\ddot{q} + b\dot{q} + kq = C_0\delta + C_1\dot{\delta} \quad [19]$$

where

$$b = -Z_\alpha - M_q - M_{\dot{\alpha}} \quad [20]$$

$$k = Z_\alpha M_q - M_\alpha \quad [21]$$

$$C_0 = Z_\delta M_\alpha - Z_\alpha M_\delta \quad [22]$$

$$C_1 = Z_\delta M_\alpha + M_\delta \quad [23]$$

It can be seen that the left-hand side of Equation [19] is the same as that for the familiar spring-mass-damping system (or the R-L-C electrical network). This equation can be expressed in a more convenient form, however, by applying the Laplace transform for zero initial conditions and solving algebraically for the ratio of the transformed quantities, q and δ. Hence

$$\frac{q}{\delta} = \frac{C_0 + C_1 s}{s^2 + bs + k} \quad [24]$$

This expression is the transfer function which defines in operational form the pitching-velocity response of the airplane to a deflection of the elevator. The replacement of s with $i\omega$ gives the frequency response directly, or the use of the inverse Laplace transform makes possible the calculation of the time response to any type of elevator motion. There is a similar transfer function for α/δ which has the same denominator but different coefficients in the numerator.

The form of the transfer function indicates an oscillatory system with a natural frequency and damping ratio which depend on the mass, inertia, and aerodynamic characteristics of the airplane. To many, a more familiar expression for the transfer function of Equation [24] may be

$$\frac{q}{\delta} = \frac{K(1 + \tau s)}{1 + \frac{2\zeta}{\omega_n}s + \frac{s^2}{\omega_n^2}} \quad [25]$$

The oscillatory motion defined by the two-degrees-of-freedom longitudinal system is termed the short-period mode and has become of greater concern as flight speeds and altitudes increase because of corresponding increases in natural frequency and decreases in damping ratio. For most airplanes, a representative range of natural frequencies is 2 to 8 radians per second, while damping ratios as low as 0.2 are not uncommon.

If the third degree of freedom (X-force equation) had been included in the derivation, the transfer function for q/δ would be of fourth order and would contain, in addition to the short-period mode, a long-period oscillatory mode termed the "phugoid." Although very lightly damped, the phugoid has such a long period (1 to 2 min) that it is usually of minor importance when considering dynamic responses.

Similar derivations can be made for the lateral system (6, 7). Here it is generally necessary to consider all three degrees of freedom, and solutions for p, r, and v to either rudder or aileron motions exhibit three modes: An oscillatory "Dutch roll" mode, a first-order "spiral" mode (often divergent), and a first-order "rolling" mode (usually well damped). For the remainder of this paper, however, we will restrict the discussion to the simplified longitudinal response characterized by Equation [24].

It is apparent now that the response of an airplane to any type of elevator-control motion can be predicted if, in addition to the mass and inertia constants, the derivatives which defined the rate of change of aerodynamic forces and moments with α, $\dot{\alpha}$, q, and δ are known. As mentioned previously, these often are not known with any degree of accuracy, particularly in a new or unconventional airplane designed to fly at or near sonic velocity. It is necessary, therefore, to resort to flight-test procedures to determine the dynamic responses and, from them, the transfer functions that are required for the adequate synthesis of automatic-control systems.

Flight Evaluation of Frequency Response

The most direct method for obtaining a frequency response of a physical system is to measure the steady-state response to a sinusoidal input at a number of different frequencies. While this procedure has been used successfully to obtain frequency responses of aircraft in flight (8), it is not feasible in many cases. This method requires the installation of equipment to drive the control surface in a sine-wave pattern and may involve an excessive amount of flight time, particularly if data are to be taken at a large number of frequencies.

In general, a more practical method of obtaining a frequency response is to analyze the transient response of the aircraft to an arbitrary deflection of the control surface (generally a step or pulse). A single transient contains the entire frequency spectrum and can be obtained in flight in a matter of seconds. A transient wave shape may be considered to be made up of a sum of sinusoidal wave shapes of various amplitudes and covering the entire frequency band. Therefore the response of a linear system to a transient input can be viewed as its response to the sum of sinusoidal waves contained in the transient input.

The basic procedure for converting transient data from the time to the frequency domain is based on the use of the Fourier integral which, under certain conditions, enables a time function $f(t)$ to be transformed into a complex frequency function $f[i\omega]$.[8] The Fourier integral may be expressed as

$$f[i\omega] = \int_0^\infty f(t)e^{-i\omega t}\,dt \quad [26]$$

This integral must be evaluated from time zero to infinity for

[8] The parentheses ( ) will be used to indicate a function of time. The brackets [ ] will denote a function of frequency

each frequency ω at which $f[i\omega]$ is desired. The integration can be accomplished only if the behavior of $f(t)$ is known for an infinite time after the disturbance is initiated. This means that it is necessary for the system to reach a steady state in some finite time T so that $f(t)$ may be expressed analytically between the time limits T and infinity. If, in the analysis of flight data, the input quantity becomes steady after a reasonably short time, the output response or one of its derivatives will also become constant within some time T, provided the system is stable and linear. Thus in the case of the pitching-velocity response to such an elevator input, the Fourier integral can be written in two parts as follows

$$q[i\omega] = \int_0^T q(t)e^{-i\omega t}\,dt + \int_T^\infty q_T e^{-i\omega t}\,dt \quad [27]$$

where q_T is the constant steady-state value of $q(t)$ corresponding to a constant value of the elevator deflection $\delta(t)$ (if a pulse input were used in which $\delta(t)$ returned precisely to zero, then q_T should also be zero). The second integral may be evaluated analytically so that

$$q[i\omega] = \int_0^T q(t)e^{-i\omega t}\,dt + \frac{q_T}{i\omega}e^{-i\omega T} \quad [28]$$

For computational purposes, it is then necessary to separate $q[i\omega]$ into its real and imaginary parts as follows

$$q[i\omega] = R + iI \quad [29]$$

where

$$R = \int_0^T q(t)\cos\omega t\,dt - \frac{q_T}{\omega}\sin\omega T \quad [30]$$

$$I = -\int_0^T q(t)\sin\omega t\,dt - \frac{q_T}{\omega}\cos\omega T \quad [31]$$

The integrals involving the transient part of $q(t)$ must then be evaluated by one of several numerical integration methods in order to obtain R and I. After this has been done, we may express $q[i\omega]$ in polar notation

$$q[i\omega] = |q|e^{i\varphi_q} \quad [32]$$

where $|q|$ is the magnitude $\sqrt{R^2 + I^2}$ and φ_q is the phase angle $\tan^{-1}(I/R)$.

This process is repeated for the input quantity $\delta(t)$ and the system frequency response is then expressed as the ratio of the two frequency functions

$$\frac{q[i\omega]}{\delta[i\omega]} = \frac{|q|e^{i\varphi_q}}{|\delta|e^{i\varphi_\delta}} = \frac{q}{\delta}e^{i(\varphi_q - \varphi_\delta)} = \frac{q}{\delta}e^{i\varphi} \quad [33]$$

The amplitude ratio $|q/\delta|$ is the ratio of the individual amplitudes, and the phase angle φ is the difference between the individual angles.

As mentioned previously, there are several practical numerical integration methods that may be used in the evaluation of the Fourier integral. One of the most common is Simpson's area rule which is used to obtain the area under the product curves $f(t)\cos\omega t$ and $f(t)\sin\omega t$ by a series of parabolic approximations. To insure reasonable accuracy with this method, the time history of $f(t)$ must be evaluated at small enough time intervals Δt so that each cycle of the product curve can be defined by at least eight measurements. This means that $\omega\Delta t$, where Δt is the time interval between measurements, should not exceed 45 deg. When considering the short-period longitudinal response of an airplane, for example, the frequency range of interest might extend from zero to 16 radians per second and thus would require a tabulation at 0.05-sec intervals.

A refinement to this basic method introduced first by Filon (9) and later applied by Schumacher (10) enables the parabolic approximation to be applied directly to the time function $f(t)$ rather than to the product curves $f(t)\cos\omega t$ and $f(t)\sin\omega t$. This allows greater accuracy in the determination of high-frequency components and it is generally possible to obtain reasonable accuracy with $\omega\Delta t$ as high as 120 deg, provided, of course, that enough time intervals are used to define the time function adequately. The application of this method is the same as for the basic Simpson's rule, except that the factors 1/3, 4/3, and 2/3 now become functions of $\omega\Delta t$, i.e., a different set of factors for each frequency at which the analysis is to be accomplished.

Other numerical integration methods based on incremental-step or triangular area approximations (11, 12) have been used successfully. Graphical methods such as developed by Seamans, Blasingame, and Clementson (12) are also applicable. However, the numerical methods based on the parabolic approximation appear to offer greatest accuracy.

These numerical methods are readily adaptable to IBM machine calculation, where the procedures are essentially the same as by hand computing. In either case the calculations are detailed and time-consuming. Probably the greatest advance in this field is the development of electronic harmonic analyzers which calculate very rapidly the frequency response directly from a graph of the time function, thus eliminating the necessity for tabulating the data. A device is currently being developed at the Ames Aeronautical Laboratory of the NACA which employs a television-type flying spot scanner to convert a graphical plot of the transient response into a voltage which is then multiplied by the $\cos\omega t$ and $\sin\omega t$ factors and integrated to give a voltage proportional to the terms

$$\int_0^T f(t)\cos\omega t\,dt$$

and

$$\int_0^T f(t)\sin\omega t\,dt$$

About 1 min is required to determine these values at each frequency. At the present this saving in time is accomplished at a sacrifice in accuracy, but it is hoped that further development will result in an improvement in this respect.

Regardless of the integration method used to obtain a frequency response, the geometric shape of the input function has a profound bearing on the frequency range through which reliable transforms can be obtained and it should be given careful consideration. The Fourier transform of the input is an indication of the excitation that is applied to the system at any frequency. Two extreme types of input are the pure step and the pure impulse.

The unit step has a transform magnitude equal to $1/\omega$, thus giving infinite excitation to the zero frequency (steady-state) component at the expense of the higher frequencies. An impulse (zero time duration), on the other hand, has a constant transform magnitude over the entire frequency spectrum. Thus it appears that for most purposes the impulse is a more desirable type of input and that the step input is effective only at very low frequencies.

In dealing with physical systems a pure impulse is impossible to attain, the nearest approach being a triangular-shaped input with a finite time base T. The transform magnitude of a triangular pulse of unit area is

$$|\delta| = \frac{8}{(\omega T)^2}\left(1 - \cos\frac{\omega T}{2}\right) \quad [34]$$

It can be seen that $|\delta|$ is a periodic function of ωT and is zero when

$\omega T = 4\pi$, 8π, and so on. At these points the transform of the output quantity also would be zero, thus making the frequency response indeterminate. Other geometric pulse shapes have somewhat similar transforms, as shown in Fig. 3, where the magnitudes

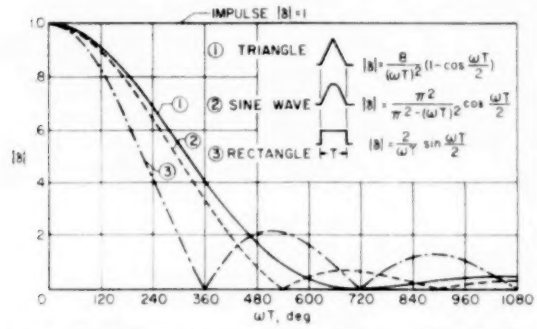


FIG. 3 TRANSFORM MAGNITUDES OF UNIT AREA PULSE INPUTS

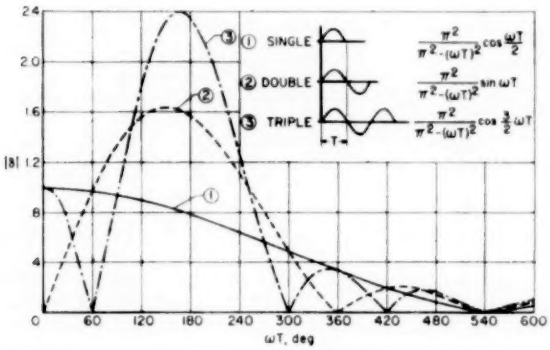


FIG. 4 TRANSFORM MAGNITUDES OF MULTIPLE SINE-WAVE PULSES

$|\delta|$ for a triangular, a half-cycle sine wave, and a rectangular pulse are plotted against the dimensionless frequency variable ωT. In each case the area under the pulse and hence the zero frequency values are unity. The three curves have similar characteristics, approach zero at values of ωT ranging from 2π to 4π, and are typical of any arbitrary-shaped pulse input.

For most accurate results, $|\delta|$ should be as large as possible over the frequency range of interest and, therefore, it is desirable to select a pulse with sufficiently small time base so that at the highest frequency ωT will be less than 360 deg. This means that when using an arbitrary pulse input with a time base of 1 sec it is generally possible to obtain reliable frequency-response data up to frequencies of about 6 radians per second. If the time base were reduced to $^1/_2$ sec the maximum reliable frequency would be increased to about 12 radians per second, provided the area under the pulse was not reduced appreciably.

In any physical system, however, there is a maximum rate at which an input may be applied. Thus to decrease T below a certain value means decreasing the area under the pulse and thereby reducing the over-all magnitude of the transform. The most desirable input, therefore, is a compromise between a large area and a short time base. However, for the case of an airplane, the control motion must be limited to sufficiently small values to assure that the system will respond in a linear manner.

Often where particularly accurate data are needed over a limited frequency range it may be desirable to use a multiple pulse as shown in Fig. 4. Here are plotted the transforms of a

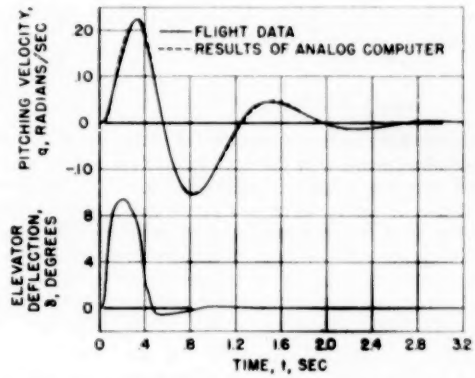


FIG. 5 TYPICAL FLIGHT RECORDS OF TRANSIENT PITCHING-VELOCITY RESPONSE TO PULSE-TYPE ELEVATOR MOTION

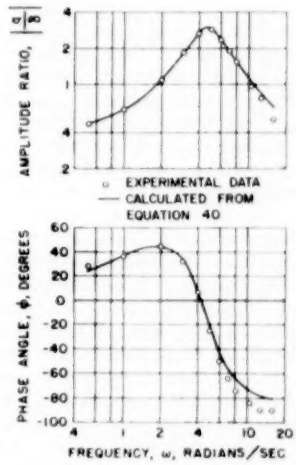


FIG. 6 PITCHING-VELOCITY RESPONSE TO ELEVATOR INPUT

single, a double, and a triple pulse. The area under each half cycle is unity. Sine-wave pulses were used in this example because of ease in calculation, but the curves are typical of any arbitrary pulse shape. Increasing the number of cycles increases the excitation at a particular frequency corresponding roughly to $\omega T = \pi$ but reduces the usable band width. For example, if an accurate transform is desired through the frequency range of 4 to 6 radians per second, then it would be desirable to use a double or triple pulse with a time base of approximately 0.6 sec.

A typical flight record of the pitching-velocity response to a pulse-shaped motion of the elevator is shown in Fig. 5. These curves are reproductions of photographic traces from recording flight instruments. Values of q and δ were tabulated at 0.05-sec intervals and analyzed by the method in which the time functions are approximated by a series of parabolas. For this sample $q(t)$ reaches a steady value of zero after about 3.5 sec and thus the second integral of Equation [27] is zero. For example, at $\omega = 4$ radians per second, it is then necessary to evaluate

$$R = \int_0^{3.5} q(t) \cos 4t \, dt = 0.0847 \qquad [35]$$

and

$$I = -\int_0^{3.5} q(t) \sin 4t \, dt = -0.0777 \qquad [36]$$

This process is repeated for $\delta(t)$, giving (at $\omega = 4$) $R = 0.0279$ and $I = -0.0331$. Then the ratio

$$\frac{q[i\omega]}{\delta[i\omega]} = \frac{0.0847 - i\,0.0777}{0.0279 - i\,0.0331} = 2.65e^{i7.36°} \quad \ldots\ldots [37]$$

This numerical integration procedure was carried out at a number of frequencies ranging from 0.5 to 16 radians per second, and the resulting frequency response is shown in Fig. 6. The amplitude curve shows, at about 4.5 radians per second, a resonant peak that is typical of second-order oscillatory systems. The location of this peak generally defines the frequency range of interest; however, as mentioned previously, the accuracy to which responses can be evaluated through this range depends on the type of input used and on the time interval at which the transient data are tabulated.

Reduction of Frequency-Response Data From Graphical to Equation Form

While the frequency response in graphical form may be sufficient for many purposes, it is often desirable or even necessary (as for analog-computer studies) to express this information in the analytical transfer-function form, particularly when considering complex systems. Numerical values for the transfer-function coefficients may be determined by operating on graphical frequency-response data, or in some cases they may be evaluated directly from the original transient-time history. In either event a curve-fitting problem is involved, i.e., a problem of finding the numerical equation that best describes a section of experimental data.

Curve-fitting methods may be divided into two general types: analytical and graphical. The analytic methods (13 to 17) often involve the use of the principle of least squares, and in all cases it is necessary to know, or at least to assume in advance, the form of the transfer function, i.e., the degree of both the numerator and denominator polynomials whose ratio is the transfer function. The problem is then to find, for a given set of experimental data and an equation of a particular form, the best numerical values of the coefficients so that the equation most nearly describes the data. Methods of this type involve a considerable amount of labor, particularly if the assumed transfer function is of higher than second order. It also is necessary to use data with a degree of accuracy not always attainable from flight tests.

Certain graphical methods offer practical advantages over the foregoing analytic procedures in the analysis of flight data. One method for obtaining numerical transfer-function coefficients from graphical frequency-response data used successfully at the Ames Aeronautical Laboratory eliminates much of the tedious calculations, allows a quick appraisal of the data (often flight data are invalid because of extraneous disturbances such as rough air), and does not require prior knowledge of the form of the transfer function.

This method consists of fitting a graphical frequency response with a set of dynamic response templates which define the first and second-order complex functions

$$\log\,[1 + i\omega\tau] \text{ and } \log\left[1 + 2\zeta i\left(\frac{\omega}{\omega_n}\right) - \left(\frac{\omega}{\omega_n}\right)^2\right]$$

where $\omega\tau$ and ω/ω_n are dimensionless frequency variables and ζ the damping ratio. These templates made from transparent plastic were developed by Dr. C. S. Draper of the Instrumentation Laboratory at the Massachusetts Institute of Technology and their use is described in a paper by Lees (18). The set contains one pair of first-order curves (log amplitude and phase angle) and a family of second-order curves with different values of the parameter ζ. Samples of these templates are shown in Figs. 7 and 8.

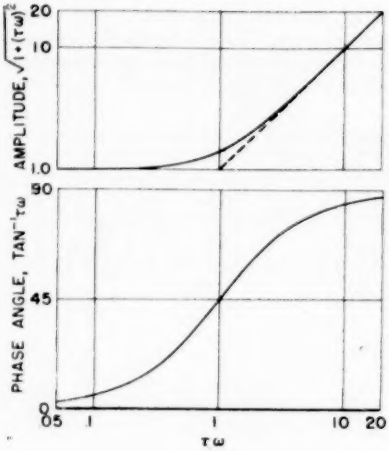


Fig. 7 (*left*) Frequency Response of First-Order Term $1 + \tau i\omega$

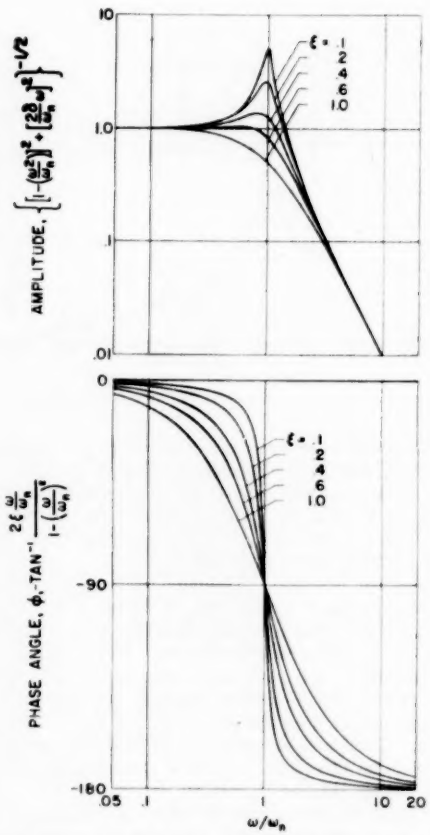


Fig. 8 Frequency Response of Second-Order Term $\left[1 + \frac{2\zeta}{\omega_n} i\omega - \left(\frac{\omega}{\omega_n}\right)^2\right]^{-1}$

The value of unity on the dimensionless frequency scales ($\omega\tau$ and ω/ω_n) is termed the "breakpoint."

In general, for a linear system, the polynomials which make up the numerator and denominator of the transfer function can be factored into first and second-order terms of the types shown in

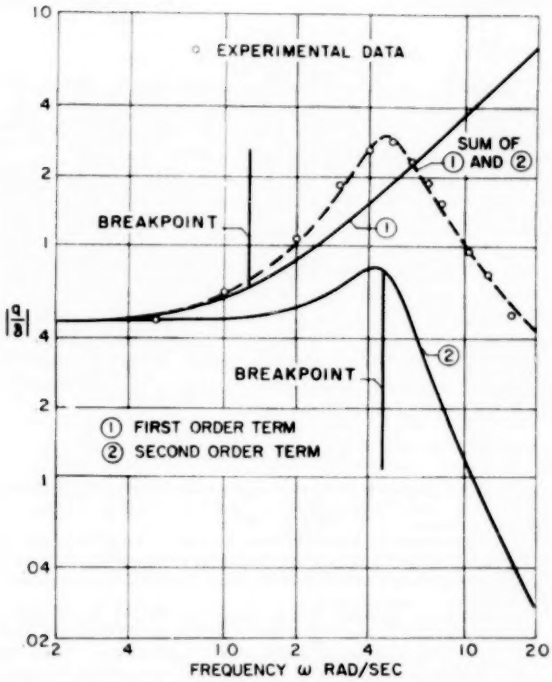


FIG. 9 MATCHING OF EXPERIMENTAL FREQUENCY RESPONSE BY MEANS OF TEMPLATES

the foregoing. The graphical frequency response can then be constructed by locating properly the necessary templates with regard to frequency and by summing up the log amplitudes and the individual phase angles.

The inverse problem of finding the first and second-order terms that make up a graphical frequency response also can be solved. The frequency-response data are plotted to the correct logarithmic scale, and by combining templates in the proper manner it is generally possible after a few trials to find one combination that will match closely both amplitude ratio and phase angle. This particular combination indicates the form of the equation which most nearly describes the graphical frequency response.

Fig. 9 illustrates the manner in which the experimental amplitude-ratio curve in Fig. 6 is matched by a first-order numerator term and a second-order denominator term with $\zeta = 0.3$. (The corresponding match of the phase-angle plot is not shown.) The breakpoint of the first-order curve occurs at $\omega = 1.27$; thus $\tau = 1/1.27 = 0.82$. Similarly, the second-order breakpoint gives $\omega_n = 4.65$. The gain constant K (the zero frequency amplitude) is 0.460. The resulting transfer function is then

$$\frac{q}{\delta} = \frac{0.460\,(1 + 0.82s)}{1 + \dfrac{2 \times 0.3}{4.65}s + \left(\dfrac{s}{4.65}\right)^2} \qquad [38]$$

The numerical transfer functions obtained with the templates may be sufficiently accurate for many purposes; however, an additional refinement of these data involving the use of an electronic analog computer (with input and output tables) is sometimes desirable. In this procedure the transfer function with numerical values as determined by the templates is set up on the computer. A record of the actual time history of the input disturbance used in flight is supplied as an input to the computer; the output is then compared with the response that had been obtained in flight. If the records match exactly, the coefficients obtained from the templates are correct. If the records do not agree, the

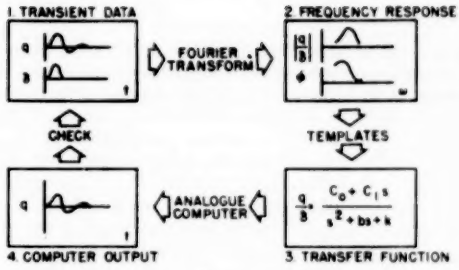


FIG. 10 DATA-REDUCTION PROCEDURES

values of the coefficients used on the computer can be modified readily until a close match is obtained. This operation not only refines the data from the templates but serves as a complete check of the entire analysis.[4] Fig. 10 is a graphical illustration of the data-reduction procedure demonstrated in this paper.

For the data used in the previous example, the transfer function obtained from the templates was modified on the analog computer until the close match shown in Fig. 5 was obtained. The corresponding final transfer function is

$$\frac{q}{\delta} = \frac{0.415(1 + 1.04s)}{1 + \dfrac{2 \times 0.356}{4.70}s + \left(\dfrac{s}{4.70}\right)^2} \qquad [39]$$

which also may be expressed as

$$\frac{q}{\delta} = \frac{9.16 + 9.52s}{s^2 + 3.35s + 22.1} \qquad [40]$$

The frequency response calculated from this equation has been plotted in Fig. 6 and shows excellent agreement with the measured response except at frequencies greater than 8 rps. The deviations at the higher frequencies may result from aeroelastic deformations of the airplane or from inaccuracies in the Fourier analysis of the transient data. In any event, the discrepancy is not reflected in the time history of Fig. 5. The variations in the coefficients of Equations [38] and [39] are indicative of the accuracy that can be attained by the use of the templates alone.

It can be seen that the experimentally evaluated transfer function is of the same form as predicted from the theoretical equations of motion (Equation [24]) and indicates that the assumptions leading to Equation [24] are valid in describing the motions of this airplane.

EFFECTS OF SPEED AND ALTITUDE VARIATIONS ON AIRPLANE RESPONSE

A transfer function obtained from flight data is valid only for the flight conditions (speed and altitude) at which the test was made. It may be desirable, however, to predict the response of the airplane at some different operating condition at which flight data are not available. The transfer-function coefficients can be extrapolated in certain cases to give the desired information.

According to basic theory, the aerodynamic forces and moments acting on an airplane are proportional to the dynamic pressure $\rho V^2/2$ and thus Z and M are commonly expressed as

$$Z = -L = -C_L \frac{\rho V^2 S}{2} \qquad [41]$$

$$M = C_m \frac{\rho V^2 S c}{2} \qquad [42]$$

[4] When dealing with familiar systems it is often possible to omit completely the frequency-response calculations and to use the analog computer to obtain transfer functions directly from transient-time histories, provided a reasonable estimate of the transfer function is available.

where L is the total lift force, and C_L and C_m are dimensionless lift and moment coefficients. Using this notation, the partial derivatives of Equations [17] and [18] may be written as

$$Z_\alpha = -\frac{\rho VS}{2m} C_{L\alpha} \qquad M_{\dot\delta} = \frac{\rho V^2 Sc}{2I_y} C_{m\dot\delta}$$

$$Z_{\dot\delta} = -\frac{\rho VS}{2m} C_{L\dot\delta} \qquad M_q = \frac{\rho VSc^2}{4I_y} C_{mq}$$

$$M_\alpha = -\frac{\rho V^2 Sc}{2I_y} C_{m\alpha} \qquad M_{\dot\alpha} = \frac{\rho VSc^2}{4I_y} C_{m\dot\alpha}$$

where $$C_{L\alpha} = \frac{\partial C_L}{\partial \alpha}, \text{ etc.}$$

It is then possible to express the coefficients b, k, C_0, and C_1 (Equation [19] after neglecting terms which have been shown by experience to be negligibly small) as

$$b = \frac{\rho VS}{2}\left[\frac{C_{L\alpha}}{m} - (C_{mq} + C_{m\dot\alpha})\frac{c^2}{2I_y}\right] \qquad [43]$$

$$k = -\frac{\rho V^2 Sc}{2I_y} C_{m\alpha} \qquad [44]$$

$$C_0 = \frac{\rho^2 V^2 S^2 c}{4mI_y}(C_{L\dot\alpha}C_{m\alpha} - C_{L\alpha}C_{m\dot\alpha}) \qquad [45]$$

$$C_1 = \frac{\rho V^2 Sc}{2I_y} C_{m\dot\delta} \qquad [46]$$

If the stability derivatives $C_{L\alpha}$, etc., remain constant over the range of flight conditions of interest, then the following proportionalities can be written directly

$$b \sim \rho V$$

$$k \sim \rho V^2$$

$$C_0 \sim \rho^2 V^2$$

$$C_1 \sim \rho V^2$$

Since the density ρ is a direct function of altitude, these relations allow the extrapolation of experimental transfer-function coefficients and the related frequency response over a range of speed and altitude.

There are certain limitations in the use of these relations which should be mentioned. While the stability derivatives are essentially independent of velocity, they do vary with Mach number. The nature of this variation depends on the particular airplane configuration but is generally appreciable only at Mach numbers greater than 0.5. Through the transonic range, however, the variations are apt to be quite large and erratic. Strictly speaking then, for high-speed airplanes, the proportionalities would then be valid only for speed and altitude changes corresponding to a constant Mach number.

There are still other limitations which must be considered. At very low speeds or at extremely high altitudes the airplane by necessity is trimmed at a large angle of attack. Extreme angles of attack may lead to drastic changes in some of the stability derivatives. Furthermore, in the case of relatively flexible airplanes, aeroelastic deformations will cause the effective values of certain derivatives to vary with dynamic pressure. In spite of these limitations, the foregoing relations are a useful guide in estimating trends that will occur with changes in speed and altitude.

Conclusion

The growing need for dynamic-response data on aircraft has opened an entirely new field in flight testing and analysis. Information of this type is of extreme importance in the design of autopilots and completely automatic control systems.

While dynamic responses may be calculated in many cases from conventional stability derivatives as obtained from wind-tunnel and theoretical studies, it is often necessary to resort to flight-testing techniques in order to obtain accurate data over the wide ranges of operating conditions of present and future aircraft. Dynamic-response data may be presented in frequency-response form but are of most general interest when expressed in analytical transfer-function form. The use of certain graphical aids in conjunction with electronic analog-computing equipment leads to rapid and reasonably accurate evaluation of airplane transfer functions from transient flight data.

BIBLIOGRAPHY

1 "Aerodynamic Stability and Automatic Control," by William Bollay (The 14th Wright Brothers Lecture), *Journal of the Aeronautical Sciences*, vol. 18, 1951, pp. 569–617.

2 "Dynamic Stability and Control Research," by W. F. Milliken, Jr., presented at 3rd International Joint Conference of the RAS-IAS, Brighton, England, September 3–14, 1951, Cornell Aero. Lab., Rep. No. Cal-39.

3 "Dynamic Analysis of Aeroelastic Aircraft by the Transfer Function-Fourier Method," by J. B. Rea, *Journal of the Aeronautical Sciences*, vol. 18, 1951, pp. 375–397.

4 "Aerodynamic Theory," by W. F. Durand, Julius Springer, Berlin, Germany, vol. 5, 1935.

5 "Airplane Performance Stability and Control," by C. D. Perkins and R. E. Hage, John Wiley and Sons, Inc., New York, N. Y., 1949.

6 "Automatic Flight Control—Analysis and Synthesis of Lateral Control Problem," by R. N. Bretoi, Trans. ASME, vol. 74, 1952, pp. 415–530.

7 "Summary of Methods for Calculating Dynamic Lateral Stability and Response and for Estimating Lateral Stability Derivatives," by J. P. Campbell and M. O. McKinney, NACA TN 2409, 1951.

8 "Dynamic Longitudinal Stability and Control Flight Tests of a B-25J Airplane—Forced Oscillation and Step Function Response Methods, Utilizing an A-12 Automatic Pilot," by G. F. Campbell, D. W. Whitcomb, and W. O. Breuhaus, Cornell Aero. Lab. Rep. No. TB-405-F-3, April, 1947 (USAF Tech. Rep. 5688).

9 "On a Quadrature Formula for Trigonometric Integrals," by L. N. G. Filon, Proceedings of the Royal Society of Edinburgh, Scotland, vol. 49 (XLIX), 1928-1929, pp. 38–47.

10 "Methods of Analyzing Transient Flight Data to Obtain Aircraft Frequency Response," by L. E. Schumacher, WADC Memo. Rep. No. MCRFT-2268, January, 1950.

11 "Application of the Performance Operator to Aircraft Automatic Control," by R. C. Seamans, Jr., B. G. Bromberg, and L. E. Payne, *Journal of the Aeronautical Sciences*, vol. 15, 1948, pp. 535–555.

12 "The Pulse Method for the Determination of Aircraft Dynamic Stability," by R. C. Seamans, Jr., B. P. Blasingame, and G. C. Clementson, *Journal of the Aeronautical Sciences*, vol. 17, 1950, pp. 22–38.

13 "A Survey of Methods for Determining Stability Parameters of an Airplane From Dynamic Flight Measurements," by H. Greenberg, NACA TN 2340, 1951.

14 "A Least Squares Curve Fitting Method With Applications to the Calculation of Stability Coefficients From Transient Response Data," by M. Shinbrot, NACA TN 2341, 1951.

15 "Matrix Method of Determining the Longitudinal-Stability Coefficients and Frequency Response of an Aircraft From Transient Flight Data," by J. J. Donegan and H. A. Pearson, NACA Rep. 1070, 1952.

16 "A Survey of Stability Analysis Techniques for Automatically Controlled Aircraft," by A. L. Jones and B. R. Briggs, NACA TN 2275, 1951.

17 "Application of Several Methods for Determining Transfer Functions and Frequency Response of Aircraft From Flight Data," by J. M. Eggleston and C. W. Mathews, NACA TN 2997, 1953.

18 "Graphical Aids for the Graphical Representation of Functions of the Imaginary Argument," by S. Lees, M.I.T. Instrumentation Lab. Eng. Memo. E-25, February, 1951.

Discussion

R. K. KOEGLER.[5] This excellent paper presents first a brief but clear and accurate review of the problem of evaluating frequency-response characteristics of aircraft from flight-test data and the history of such work. This is followed by a clear presentation of a template-fitting method for obtaining numerical transfer-function coefficients from graphical frequency-response data and a method for extrapolating transfer-function expressions to flight conditions other than those at which the test data were obtained.

As indicated by the authors, it is necessary to have the dynamic stability response characteristics of an aircraft in the form of transfer-function coefficients for many calculation purposes; and it is not possible to calculate these coefficients unless the form of the equations of motion is well known. Also, as pointed out, such factors as aeroelasticity and rapid changes in aerodynamic characteristics with Mach number lead to confusion and uncertainty regarding the equations to be used. The method presented of obtaining transfer-function expressions from flight-test data on frequency-response plots circumvents this difficulty.

Our experiences in this type of work lead us to see further advantages of this method in another slightly different direction. For design purposes, recourse to flight-test data is not possible. Therefore the designer needs the best possible information on the stability derivatives, the methods of their estimation, and the equations of motion with which they should be used. The method presented in this paper should be very helpful in determining the transfer-function expressions and hence equations of motion to be used for various existing aircraft, and this experience can be utilized in design calculations. Such experience will be helpful not only where aeroelastic effects, transonic Mach-number effects, and the like, are involved, but also in work on helicopters and in cases where the modes of motion are not well separated. In addition, information on the transfer-function expressions or equations of motion are necessary for the accurate determination of the stability derivatives from flight-test data. This is true whether these derivatives are determined by simultaneous solutions of the equations of motion or by an extension to this problem of the extrapolation procedure presented in this paper.

SIDNEY LEES.[6] The authors have presented an excellent compact review of the pulse method for testing systems. This is an important method which has been developed in recent years to determine and test the dynamic properties of systems. The paper may be divided into two parts, one relating to the pulse method of testing and the other to determining the dynamic properties of aircraft. The more generally useful is the description of the pulse method.

At the M.I.T. Instrumentation Laboratory, the method has been used not only for aircraft, but for automatic-control systems including fire-control systems. Prof. Joel Hougen, at Rensselaer Polytechnic Institute, has been investigating pulse testing in chemical-engineering applications such as heat exchangers. Other applications are being developed.

The general background for pulse testing is given by Draper, McKay, and Lees.[7] In particular, the mathematical basis for numerical Fourier transforms and graphical forms of transfer functions is developed in chapter 25 of that text.

In the paper the authors mention two extreme types of inputs, namely, a pure step function and a pure impulse. The pure impulse is Fourier transformable but a pure step function is not. This is a mathematical objection but it means that the numerical integration methods will not produce a finite answer. In order to overcome this difficulty, both the step-function input and its response are numerically differentiated before transformation. However, the derivative of a step function is an impulse. The derivative of a step-function response is an impulse-function response. Mathematically, the frequency spectrum for an impulse-function input and a step-function input may be made identical by taking derivatives. The main difficulty is that numerical processes of differentiation lack precision. As the authors imply, there is a choice of the form of the input function. In some instances the use of step-function inputs is unavoidable.

It is not obvious why the frequency-response data are limited so that ωT is less than 360 deg. This may depend upon the convergence of the Fourier integral given by Equation [27] under numerical analysis. The convergence of numerical processes depends upon several factors such as the number of points employed to represent the time function and the type of time-function approximation.

A word concerning the reduction of the frequency-response data to equation form by use of templates. As the authors state, it is helpful to find an expression with constant coefficients for a frequency-function gain in graphical form. Mathematically, this is an approximation of the graphical expression by a linear differential equation with constant coefficients. One form of the differential equation is the transfer function consisting of a ratio of two constant-coefficient polynomials. Both the numerator and denominator polynomials are expressible as products of first and second-order factors with real coefficients.

Conventionally, approximations of graphical functions are expressed as a series of orthogonal functions like a Fourier series. The approximations are made in a straightforward manner and produce a unique expression. The transfer-function type of expression is a product function, not a series.

Its factors are not orthogonal and, most important, the approximation is not unique. The chief virtue of the transfer-function approximation is that the form is most familiar and most useful to an engineer. By using the log of the transfer function, the product becomes the sum of the log factors each of which is either a first or a second-order factor. However, the terms of the log transfer function are nonorthogonal and the approximation is still not unique. The templates referred to in the paper represent the log factors. The approximation is carried out by graphical trial and error for the best fit. The mean-square error can be estimated. This is a practical method which was first explored by R. T. Donohue. A discussion of the log frequency method including the templates is given in chapter 24 of the reference cited.[7]

The authors describe a very nice method of refinement of the approximation by using an analog computer. This further emphasizes the experimental nature of the approximation. It is not a method that would be used by mathematicians. Its basic appeal is to engineers.

RUFUS OLDENBURGER.[8] Equilibrium drift and the amount of time required for experimental runs are often serious considerations in frequency-response work. As pointed out by William R. Ahrendt, this is particularly true of the domain of process control as well as that of the automatic piloting of aircraft. On the other hand, the time for converting from transient to frequency-response data is by no means negligible. At the writer's company direct frequency-response runs are preferred.

[5] Flight Research Department, Cornell Aeronautical Laboratory, Inc., Buffalo, N. Y.

[6] Assistant Professor, Aeronautical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[7] "Instrument Engineering," by C. S. Draper, W. McKay, and S. Lees, vol. 2, McGraw-Hill Book Company, New York, N. Y., 1953.

[8] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

N. L. Wener,[9] D. Graham,[9] R. C. Lathrop.[9] The authors have presented a clear and concise exposition of the mathematical background for the measurement of aircraft dynamic response. They have ably elucidated the present urgent necessity for obtaining such data as a means of verifying aerodynamic and structural theories and for actual design purposes in connection with air vehicle automatic controls. It is interesting to note that the overwhelming proportion of the directly pertinent references in the bibliography are of very recent date. It is true that the techniques discussed have been given considerable emphasis only in recent years, although the basic mathematics involved was formulated long ago by Fourier. This fact is attributable to the recent increase in the availability of data transcription and automatic computing equipment.

The analytical transfer function is not necessarily the ultimate expression of dynamic response. It is often necessary or desirable to extract the stability derivatives. This is generally more difficult than obtaining the transfer-function coefficients. Additional investigation into methods of extracting the stability derivatives from flight data is sorely needed, particularly for the higher-order systems.

The residual response indicated by Equations [27] through [31] for $T < t < \infty$ is only one possibility. Other common types of responses can be expressed by an oscillation

$$q(t) = C \sin(\omega_0 t + \psi), \quad T < t < \infty$$

an exponential

$$q(t) = Ce^{at}, \quad T < t < \infty$$

or a ramp

$$q(t) = Ct, \qquad T < t < \infty$$

Combinations of these terminal or residual responses are also possible. Representing the response after time T by these functions is quite reasonable when the input has reached a steady state prior to T. These residual responses are transformable in the same manner as the constant residual response of Equations [27] and [28].

For several years the All-Weather Branch, Wright Air Development Center, has utilized punch-card calculators for obtaining frequency response from transient response. For testing large quantities of data, this method has proved to be the best yet investigated. Without sacrificing accuracy, it appears to exceed the speed of the flying-spot function-generator method described by the authors. This latter method also was investigated by the writers. It is more versatile than the card-programmed method, in that any frequency can be evaluated, not just integral harmonics. However, the time required to prepare a copy of the transient to be analyzed, the difficulty in handling residual response, and necessity of converting from complex to polar co-ordinates added up to slower operation than was obtainable by the card-programmed computer technique. These are not all inherent limitations, and presumably could be overcome. We plan to conduct further investigations into analog methods, using an automatic curve follower to introduce the transient function. We also have utilized a digital differential analyzer in a method similar to the analog method, and, except for improved accuracy, with the same general results.

The punch-card computer method does require certain equipment for effective use. An oscillogram reader, a storage counter, and a card punch provide a means for transcribing rapidly flight-test recordings into digital form. When operated on a production basis, the transcription to punched cards can be accomplished in approximately 10 min per trace (90–120 samples). The computation, including conversion to polar co-ordinates, requires approximately 25 min to obtain 30 frequencies from 90 time samples. It is relatively simple to obtain amplitude ratio and phase angle on the punch-card computer, since the necessary input and output time histories are already available in punched-card form.

One variation of a least-squares method of computing transfer-function coefficients (or stability derivatives) is given elsewhere.[10] This method has been carried out automatically on punched-card calculators, thus eliminating the tedious hand calculations otherwise required. Thus far the method has been applied by the writers to the longitudinal mode only, using a second-order system for n/δ, where n is the acceleration along the Z-axis. The authors' comment regarding the need for data having accuracy not always obtainable from flight test finds support here. Even the best of flight data, obtained by steady-state oscillation tests under ideal conditions, produce less than optimum results when applied to the determination of transfer-function coefficients and stability derivatives. The combined use of flight-data reduction and the analog computer to match transient response seems to be a good approach to the problem.

[9] All-Weather Branch, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio.

Authors' Closure

The authors wish to express their appreciation for the comments of the reviewers. Mr. Koegler's comments regarding the desirability of further refinement of the methods for evaluating stability derivatives from frequency-response data are particularly true for operation of aircraft where aeroelastic and Mach-number effects are important. The development of extremely rapid data-reduction systems such as outlined by Messrs. Wener, Graham, and Lathrop certainly make pulse methods more attractive. The authors agree with Professor Lees that the recent publication of volume 2 of *Instrument Engineering* makes available an authoritative reference for pulse methods. Unfortunately, this volume was not available to the authors at the time this paper was prepared so the prior reference (18) to Professor Lees' paper was cited.

In regard to the technical comments by Professor Lees, it is agreed that in the strict mathematical sense a step function is not Fourier transformable. However, this difficulty can be circumvented without resorting to differentiation if the Fourier transform of $f(t)$ is defined as

$$f[i\omega] = \lim_{\sigma \to 0} \int_0^\infty e^{-(\sigma + i\omega)t} f(t)dt$$

It should be noted that this generalized definition is identical to the usual definition provided that the ordinary Fourier transform exists and is continuous; i.e., if

$$\int_0^\infty e^{-i\omega t} f(t)dt$$

exists, then

$$\int_0^\infty e^{-i\omega t} f(t)dt = \lim_{\sigma \to 0} \int_0^\infty e^{-(\sigma + i\omega)t} f(t)dt$$

With this definition of the Fourier transform if $\delta(t)$ is a unit step input (at $t = 0$)

$$\delta[i\omega] = \lim_{\sigma \to 0} \int_0^\infty e^{-(\sigma + i\omega)t} dt = \frac{1}{i\omega} = \frac{1}{\omega} e^{-\frac{\pi}{2}i}$$

[10] "A Method for Evaluating Aircraft Stability Parameters From Flight Test Data," by L. E. Schumacher, Wright Air Development Center, Technical Report 52-71, June, 1952.

Similarity if the response $q(t)$ to a unit step has a constant value q_T, for $T < t < \infty$; Equations [27] to [31] are valid according to the above definition. The first integral of Equation [27] is evaluated numerically while the second integral is evaluated analytically as shown to give

$$\frac{q_T}{i\omega} e^{-i\omega t}$$

Professor Lees also mentions that the reason for restricting ωT to 360 deg has not been made clear. T in this case is the time duration of a pulse input. As indicated in Fig. 3, to obtain the greatest accuracy with a pulse-type input, ωT should be as small as possible. Furthermore, Fig. 3 shows that if ωT is less than 360 deg throughout the frequency range of interest, the transform will always be greater than zero and indeterminate points will be avoided.

AN ASME PAPER

Its Preparation, Submission and Publication, and Presentation

To a large degree the papers prepared and presented under the ASME sponsorship are evidence by which its professional standing and leadership are judged. It follows, therefore, that to qualify for ASME sponsorship, a paper must not only present suitable subject matter, but it must be well written and conform to recognized standards of good English and literary style.

The pamphlet on "AN ASME PAPER" is designed to aid authors in meeting these requirements and to acquaint them with rules of the Society relating to the preparation and submission of manuscripts and accompanying illustrations. It also includes suggestions for the presentation of papers before Society meetings.

CONTENTS

PREPARATION OF A PAPER—

General Information—Style, Preferred Spelling, Length Limitation, Approvals and Clearances.

Contents of the Paper—Title, Author's Name, Abstract, Body of Paper, Appendixes, Acknowledgments, Bibliographies, Tables, Captions, Photographs, Other Illustrations.

Writing the Paper—Outline Tabulations, Tables, Graphs, Charts for Computation, Drawings, Mathematics, Accuracy, Headings and Numbering, Lantern Slides, Motion Pictures, Typing, Number of Copies.

SUBMISSION AND PUBLICATION OF A PAPER—

Intention to Submit Paper Required in Advance, Meeting Dates, Due Dates for Manuscript, Discussers, Review and Acceptance, Proofs, Advance Copies and Reprints, Discussion and Closure, Publication by Others.

PRESENTATION OF A PAPER—

Time Limit, Addressing Your Audience, Public Address Systems, Use of Slides.

REFERENCES—

References on Writing and Speaking, Engineering Standards.

Price 50¢. 20% Discount to ASME Members.

THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS

29 West 39th Street, New York 18, N. Y.

ASME PUBLICATIONS
On Control Systems and Instrumentation

DYNAMICS OF AUTOMATIC CONTROLS

A book for both the practical and theoretical worker, it discusses the dynamic properties which all controllers possess in common, freed from constructional and technological detail, and provides information on the functional elements of a control and the mathematical methods for handling control problems. It considers the important components of the control loop and special control problems. The causes of nonlinearities in the control loop and the on-off regulator and its field of application are studied, also the theory of the excursion-dependent periodic controller. Appendixes list formulas for the Laplace transformation, for difference equations, and for typical elements and combinations of the dynamical circuit.

1948 $6.00

PRINCIPLES OF OPTIMALIZING CONTROL SYSTEMS AND AN APPLICATION TO THE INTERNAL COMBUSTION ENGINE

This Report outlines the background theory of optimalizing control in which an automatic unit is placed in the feedback branch of a system to function in the manner of a human operator in searching out and holding the best performance from a controlled system, in spite of any reasonable change of the output level or environmental operating conditions. Optimalizing control systems considered are the sensitivity signal type, continuous test signal type, output sampling type, and peak holding type. Additionally, there is a discussion of the recovery time to be expected from the use of the various schemes for adjusting the control input. The engine used for an experimental check on the validity of optimalizing control theory was a single-cylinder internal combustion engine.

1951 $2.00

HIGH PRESSURE MEASUREMENT

The fourteen treatises in this book provide first-hand information on the following subjects. Piston ring as a precision measuring instrument, a high-pressure wire gage using gold-chrome wire, techniques and equipment for generation of dynamic high pressures, an application of high pressure in geophysics, design of high pressure pumps suitable for pilot plant service, an automatic sensitive pressure controller, influence of residual stress on behavior of thick wall closed-end cylinders, rupture-disk design evaluation, gasket for high pressure vessels, controlled volume pumps for pressures in range of 10,000 to 30,000 psi, seals to minimize leakage at high pressures, and a static dynamic load machine for high pressure.

1953 $3.00

AUTOMATIC CONTROL TERMINOLOGY

This 1954 Edition has been broadened to include automatic control combinations and the concepts common to the field of feedback control systems as they apply in industrial process control. Terms and their definitions are grouped under the following classifications:

Automatic Controllers; Basic Characteristics; Processes, Their Elements and Characteristics; Characteristics of Automatic Control; Types of Automatic Controller Action; Adjustments of Automatic Controller Action; Elements and Characteristics of Automatic Controllers; Basic Terms; Block Diagrams, Symbols, and Automatic Control Combinations.

Forty-six illustrative diagrams have been included to assist users to gain an understanding of the definitions.

1954 $1.00

GERMAN-ENGLISH DICTIONARY OF AUTOMATIC CONTROL TERMS (Paper No. 54—IRD-4)

This compilation of approximately 750 German words and their English translations or equivalents includes many terms which are not to be found in current German-English dictionaries. Symbols and conventions specified in the German Automatic Control Terminology Standards are also listed in order to clarify the meaning of equations and mathematical expressions found in German articles.

1954 50¢

BIBLIOGRAPHY ON BOURDON TUBES AND BOURDON TUBE GAGES (Paper No. 53—IRD-1)

This bibliography includes 143 references to articles and books on bourdon tubes and pressure gages published in the period 1849 to 1952. The bibliography is annotated, chronologically arranged, and includes an author index.

1953 50¢

THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS
29 West 39 St. New York 18, N. Y.

20% Discount to ASME Members